

Predicting Solid State Material Platforms for Quantum Technologies

Oliver Lerstøl Hebnes

*Sopra Steria, Information Technology and Services, N-4020 Stavanger, Norway and
Department of Physics and Center for Computing in Science Education, University of Oslo, N-0316 Oslo, Norway*

Marianne Etzelmüller Bathen*

Advanced Power Semiconductor Laboratory, ETH Zürich, 8092 Zürich, Switzerland

Øyvind Sigmundson Schøyen

Department of Physics and Center for Computing in Science Education, University of Oslo, N-0316 Oslo, Norway

Sebastian G. Winther-Larsen

*Menon Economics, N-0369 Oslo, Norway and
Department of Physics and Center for Computing in Science Education, University of Oslo, N-0316 Oslo, Norway*

Lasse Vines

Department of Physics and Center for Materials Science and Nanotechnology, University of Oslo, N-0316 Oslo, Norway

Morten Hjorth-Jensen

*Department of Physics and Astronomy and Facility for Rare Ion Beams,
Michigan State University, East Lansing, MI 48824, USA and
Department of Physics and Center for Computing in Science Education, University of Oslo, N-0316 Oslo, Norway*

Semiconductor materials provide a compelling platform for quantum technologies (QT), and the properties of a vast amount of materials can be found in databases containing information from both experimental and theoretical explorations. However, searching these databases to find promising candidate materials for quantum technology applications is a major challenge. Therefore, we have developed a framework for the automated discovery of semiconductor host platforms for QT using material informatics and machine learning methods, resulting in a dataset consisting of over 25.000 materials and nearly 5000 physics-informed features. Three approaches were devised, named the Ferrenti, extended Ferrenti and the empirical approach, to label data for the supervised machine learning (ML) methods logistic regression, decision trees, random forests and gradient boosting. We find that of the three, the empirical approach relying exclusively on findings from the literature predicted substantially fewer candidates than the other two approaches with a clear distinction between suitable and unsuitable candidates when comparing the two largest eigenvalues in the covariance matrix. In contrast to expectations from the literature and that found for the Ferrenti and extended Ferrenti approaches focusing on band gap and ionic character, the ML methods from the empirical approach highlighted features related to symmetry and crystal structure, including bond length, orientation and radial distribution, as influential when predicting a material as suitable for QT. All three approaches and all four ML methods agreed on a subset of 47 eligible candidates of 8 elemental, 29 binary, and 10 tertiary compounds, and provide a basis for further material explorations towards quantum technology.

INTRODUCTION

Quantum technologies (QT) based on solid state platforms have attracted a lot of attention during the last few years. Among the promising applications that are already available we find important breakthroughs such as in vivo sensing of magnetic fields in cells [1], secure communication over large distances by separation of entangled photons [2] and, notably, various quantum information processing prototypes and architectures [3]. Quantum computers are in high demand to meet the increasing need for computing power to solve complex and high-dimensional scientific problems. Recent advances have highlighted the potential of quantum information processors to outperform state-of-the-art high-performance computing facilities. Indeed, Google's 53 qubit quantum computer based on superconducting electronics solved a computational problem that was beyond the capabilities of a 200.000 core supercomputer [3]. Most recently,

* bathen@aps.ee.ethz.ch

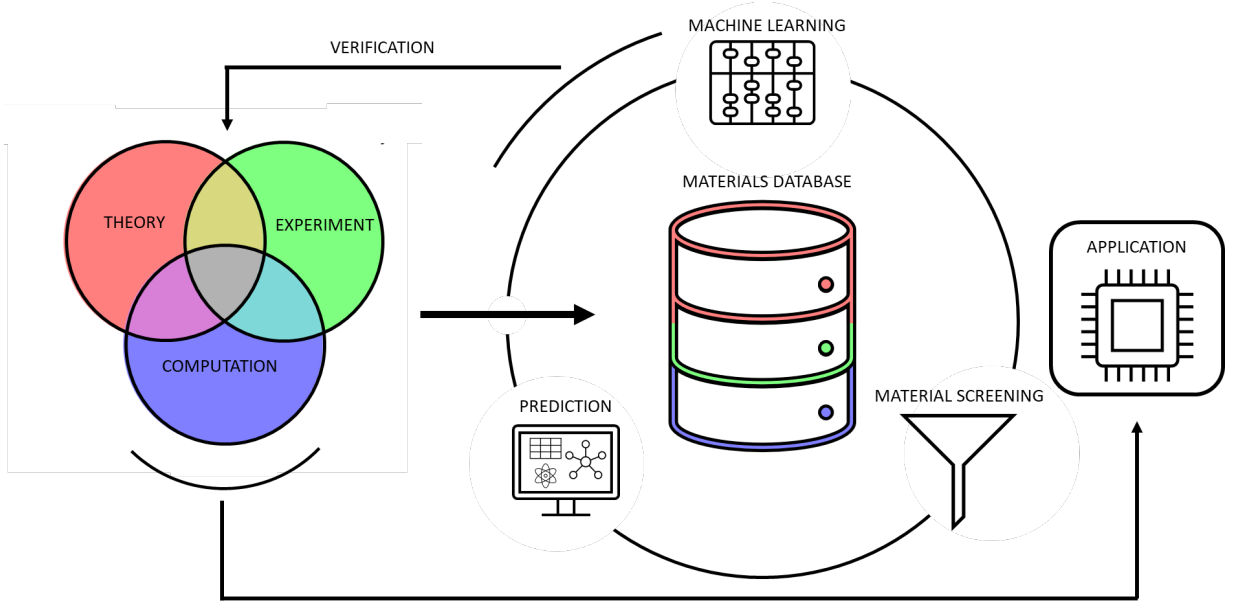


FIG. 1: Schematic of an example workflow in material informatics. Results from theory, experiment and computation are fed into material databases (arrow pointing to the center). A cycle involving material screening, machine learning and predictions leads to knowledge gain and ultimately applications in fields such as clean energy and quantum technology.

IBM announced its 127 qubit quantum processor [4]. Simultaneously, the concepts of entanglement and teleportation may eventually facilitate advanced quantum communication protocols such as quantum cryptography and the quantum internet [5], spurring further investigations into technologies based on quantum mechanics.

Several platforms are available for the development of quantum technologies, but the materials and fabrication technologies are less mature than those for, e.g., classical computers and sensors. An important concern in this context is that of scalability. For example, the best performing quantum computer prototypes available today rely on superconducting electronics that require millikelvin temperatures to operate, with the stability of interactions between qubits being an important issue. Instead, semiconductors are emerging as a promising alternative platform, offering competitive characteristics combined with the possibility of room temperature operation and mature and scalable material processing and fabrication.

Quantum technologies based on semiconductors rely on either defects or quantum dots where the latter can be of the self-assembled or nanostructured type [6]. Semiconductor defects can act as single-photon emitters or spin centers and are compatible with the three main QT categories of computing, communication and sensing [7]. These characteristics are most often found for the case of defects that introduce deep energy levels into the semiconductor band gap [8]. So-called deep level defects can trap charge carriers in localized states that are essentially isolated from the surroundings, making them highly suitable for QT due to, e.g., indistinguishable single-photon emission and long spin coherence times. The most well-known quantum compatible defect is the negatively charged nitrogen-vacancy (NV) center in diamond [9], but silicon carbide (SiC) and the various quantum emitters therein are strong contenders for quantum communication purposes especially due to the favorable emission wavelength region in the near infrared coupled with more mature material processing and fabrication (see, e.g., Refs. [10–12]). However, semiconductor based QT is still in the early stages, and the issues left to address include identification of suitable host materials and candidate defects, and scalable and reproducible quantum device fabrication. Furthermore, we lack a complete understanding of the requirements for a semiconductor material to manifest quantum compatible properties, and the selection of known quantum compatible host materials is slim [13, 14].

The majority of discoveries of QT compatible characteristics in semiconductors has so far happened by serendipity, and there is an urgent need for a better and more systematic understanding of which material requirements must be met for QT compatible characteristics like single-photon emission and single spin control to manifest. In this context, a framework for dedicated materials search and analysis is needed.

The fourth science paradigm of big data driven science reveals the potential of targeted search for promising material systems in which to expect QT compatible properties. Rather than searching through a host of signals for those that match our criteria, we aim to *predict* which materials and signatures we should target for more detailed studies, following the framework illustrated in Fig. 1. This is made feasible by the availability of databases

containing material properties for a wide range of different systems. In this work, the data in question are provided by bulk density functional theory (DFT) calculations to obtain the ground state properties of different elements and compounds. Combined with machine learning (ML) methods we provide a path towards precise classification of candidate materials. The inclusion of ML methods follows recent trends in applications of statistical learning, data science and machine learning for scientific discoveries, see for example Refs. [15, 16].

Here we provide a framework for the data mining and automated discovery of promising solid-state semiconductor hosts for quantum emitters and spin centers using targeted database search and ML methods combined with knowledge from the field. Analyzing the output of the ML methods reveals that, given a suitable initial set of labeled materials for training and testing, it is possible to discern the physical mechanisms that govern a material’s suitability for quantum applications. In this framework we do not distinguish between the specific mechanism giving rise to properties such as single-photon emission and long spin coherence times (e.g., semiconductor defects or quantum dots); instead, we attempt to target all materials that may accommodate the desired characteristics. The methodology developed herein can be modified for other material types and application areas provided that high quality databases containing relevant theoretical and/or experimental data are available.

The developed procedure relies on data extraction from the databases Materials Project [17, 18], the Open Quantum Materials Database (OQMD) [19, 20], JARVIS-DFT [21], AFLOW [22–24] and AFLOW-ML [25]. The Matminer Python library for data mining [26] was then used for material analysis to featurize the extracted data. An important aspect of our work is the database building and pertinent development of the datasets for the ML methods. We have developed three different approaches to data mining: (i) *the Ferrenti approach* which is similar to that proposed by Ferrenti *et al.* [27], (ii) *the extended Ferrenti approach* and (iii) *the empirical approach*. We find that the two first methodologies base their data extraction protocols on broad material descriptors, leading to large sets of potentially suitable candidates [28, 29]. The empirical approach, on the other hand, relies on including materials with experimentally proven advantageous characteristics in the training set and therefore yields a narrower set of possible candidates. The three resulting sets of labeled data were then analyzed with the four supervised ML methods logistic regression, decision trees, random forests and gradient boosting [29, 30], yielding 214 predicted candidates that were common between all approaches and ML methods. Example materials include ZnGeP_2 , CdS , BP , BC_2N , GeC and InP . Focused theoretical and experimental studies, using methods such as, e.g., density functional theory, photoluminescence spectroscopy and optically detected magnetic resonance, are needed to verify the predictions that quantum properties may manifest as a result of defects or nanostructures in the above listed materials. Importantly, our findings also reveal which material properties are weighted by the ML methods upon predicting a material as suitable for QT applications, thereby opening up for new discoveries in the field of quantum technologies.

RESULTS

Information Flow

The information stream in this work can be regarded as many inter-connected modular parts. The initial step for gathering material data and building features is visualized by the outer flowchart in Fig. 2 (Jupyter notebooks containing the full workflow can be found at [31]). We start by extracting all entries in the Materials Project (MP) database [17, 18] that match a specific query. The MP database contains ground state properties of different materials that are computed using density functional theory. The DFT calculations in the database were performed using the Vienna *ab initio* simulation package (VASP) [32] and the Perdew-Burke-Ernzerhof (PBE) [33] exchange-correlation functional to calculate the electronic structure of the materials. We note that despite being immensely successful in describing a number of material properties, PBE is widely known for underestimating the band gap of semiconductors [34]. Therefore, not all properties predicted using the PBE functional are reliable, and the band gaps in particular cannot be trusted in absolute terms. Nonetheless, the functional is in wide use due to the combination of reasonable accuracy and high computational throughput, and is usually considered to be reliable for large-scale trends in semiconductor properties.

The conditions for the initial MP query are that the materials must derive from the Inorganic Crystal Structure Database (ICSD) and have a band gap wider than 0.1 eV to exclude metallic compounds. The ICSD is the world’s largest database for completely identified inorganic crystal structures [36, 37]. In a parallel step, entries that are deemed similar to the entries from the initial query are extracted from the databases OQMD [19, 20], JARVIS-DFT [21], AFLOW [22–24], AFLOW-ML [25] and the Citrination platform [38]. The results of these steps are combined into a dataset for further analysis.

After material extraction we apply tools from the open-source library Matminer [26] to generate thousands of features from the data. We will refer to this process as featurization. A schematic visualization of the featurization process in Matminer is shown in Fig. 2 and focuses on a material’s composition, structure, atomic sites, density of

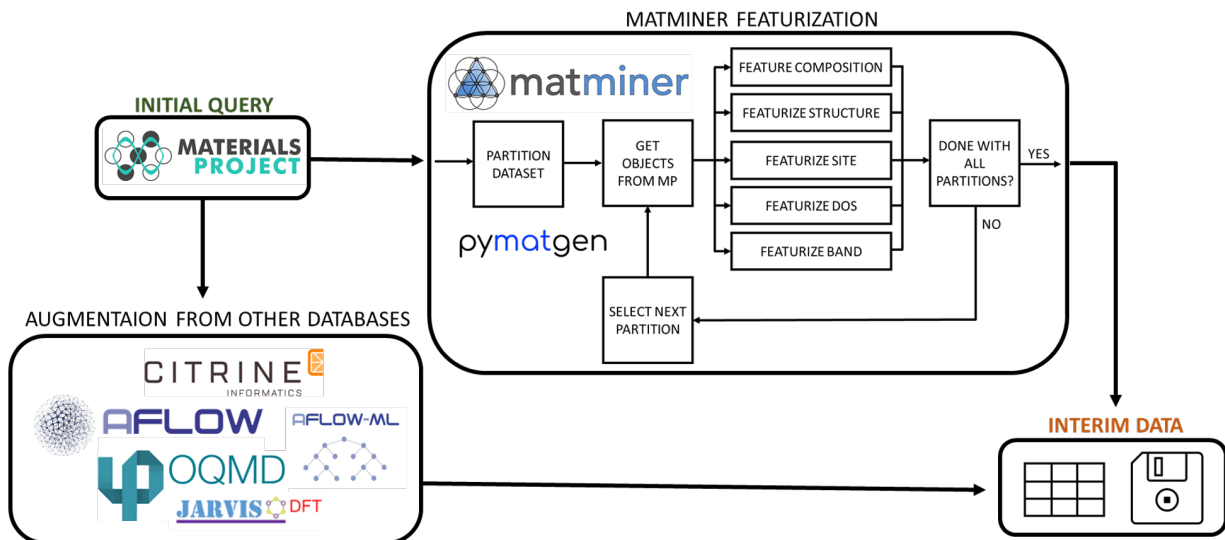


FIG. 2: The project workflow starting from an initial Materials Project (MP) query, and ending with a featurized dataset with entries from several other databases. To limit the memory and computational usage, the data are partitioned into smaller subsets where the respective Pymatgen objects (Pymatgen is a robust, open-source Python library for materials analysis [35]) are obtained through a query to be used in the following featurization steps. This process is repeated iteratively until all the data has been featurized. *DOS* refers to density of states and *band* to the electronic band structure.

states and band structure. The 39 featurizers (each generates several features) selected as material descriptors in this work are summarized and described in the Supplementary Information at [39]. The selection of features was kept rather wide as we do not want to make *ex ante* assumptions on which features best describe a solid state material platform for quantum technology.

The constructed dataset encompasses compounds formed by a plethora of combinations of surfaces, interfaces, nanostructures, compositions and structures, but this complexity is not necessarily reflected in the material descriptors. Furthermore, we have utilized data obtained from high-throughput density functional theory calculations. Indeed, there are possible errors associated with every step, starting from an initial calculation, adding of data in the database, gathering of data, featurization of data, preprocessing of data, data mining, and finally training a model and making a prediction. Unfortunately, if an error has happened in the first part of the process, the error follows the entire process and will get increasingly harder to detect. Thus, our results depend on the quality of data in the databases we employ.

Data Mining

The complete dataset consists of 25.000 materials. A subset of these materials is labeled into either suitable or unsuitable candidates for QT while the remainder will stay unlabeled. The labeled data are then grouped into training and test sets for the ML methods.

Several challenges accompany the labeling process of labeling materials according to their suitability for QT. Although QT compatible properties are becoming increasingly well studied for the case of, e.g., nitrogen-vacancy centers in diamond, single-photon emitters in silicon carbide and quantum dot (QD) structures [6, 9, 12], relatively few candidate materials are known to be suitable [13, 14]. An additional consideration is that the physical mechanisms promoting favorable properties are not fully understood. Conversely, defining materials as unsuitable candidates for QT is in many ways equally difficult, as the mechanisms preventing quantum compatible characteristics to manifest are also unknown. The strategy for selection of unsuitable candidates is thus that the negation of the criteria we use to select suitable candidates should give unsuitable candidates. A side-effect of this selection is that the criteria for unsuitable candidates can become equally, or more, restrictive compared to the criteria for suitable candidates. We will then find ourselves in a situation with skewed data sets where we have fewer unsuitable than suitable candidates. This method for selecting unsuitable candidates has shortcomings but will serve as a starting point and demonstration for how the labeling procedure could be improved. Below we follow three separate procedures for labeling materials

as suitable or unsuitable candidates for QT.

The Ferrenti approach

The first approach to labeling a selection of materials in the dataset is based on the criteria proposed by Ferrenti *et al.* [27]. They suggest a data mining process consisting of four stages by systematically evaluating the suitability of host materials taken from the Materials Project. In this framework, we label **suitable** candidates by the following steps:

1. Include materials that;
 - contain elements with more than 50 % natural abundance of spin zero isotopes,
 - crystallize in non-polar space groups,
 - are present in the ICSD database, and
 - are calculated to be nonmagnetic.
2. Pragmatically remove toxic, radioactive and otherwise “difficult” materials;
 - exclude Th, U, Cd and Hg because they are radioactive and/or toxic in the most stable forms,
 - exclude any rare-earth metals (because of the difficulty of obtaining pure materials free of isotopes with nuclear spin) and noble gases (due to the lack of stable solid phases),
 - exclude transition metal elements with unpaired electrons like Fe and Ni because of their paramagnetism; Ru and Os are also excluded because they only exist in the dataset as complex cluster structures.
3. Include only materials with a band gap larger than 0.5 eV calculated using DFT and the PBE functional. The value of 0.5 eV was chosen to match that typically predicted for silicon by PBE-level DFT calculations.
4. Ensure that the energy above hull is less than 0.2 eV per atom.

These inclusion criteria are based on the work of Weber *et al.* [8] and targeted primarily at semiconductors that can host deep level defects with spin qubit capabilities. In this context, long spin coherence times are needed, and the possibility of having an environment free of nuclear spins and permanent magnetism is therefore desirable. Moreover, non-polar materials are assumed to be preferable to obtain defect based single-photon emitters with sharp zero-phonon lines and indistinguishable single-photon emission. Transition metal elements are eliminated if they have unpaired electrons because the presence of permanent electric dipole moments may have a detrimental impact on the optical coherence of defect emission. Finally, the energy above hull requirement is included to ensure that the selected compounds are thermodynamically stable.

Next, **unsuitable** candidates are labeled according to the reverse requirements of the above; as materials in the ICSD database [36, 37] that crystallize in polar space groups, are calculated to be magnetic and have a band gap larger than 0.1 eV in the MP database (to exclude metals but include lower-band gap semiconductors). The resulting set of labeled materials contains 1581 materials where 35 % are labeled as unsuitable and 65 % as suitable for QT applications. Here we see the first example of how the criteria for labeling unsuitable candidates are more restrictive than those for the suitable candidates.

The extended Ferrenti approach

Here, we adjust the Ferrenti approach to expand the data labeling process beyond practical considerations. The second approach is therefore named *the extended Ferrenti approach* and involves removing stage two from the approach above. Moreover, we include a few additional elements that have shown promising properties but were initially excluded due to the lack of spin zero isotopes.

The following steps constitute the process of labeling **suitable** candidates in the *extended Ferrenti approach*:

1. Include materials that;
 - contain elements where more than half have a natural abundance of spin zero isotopes, including Al, P, Ga, As, B and N,
 - crystallize in non-polar space groups,

- are present in the ICSD,
 - are calculated to be nonmagnetic.
2. Only keep materials that have a band gap larger than 1.5 eV in the MP database. The higher band gap requirement (as compared to the Ferrenti approach) is included here to avoid labeling an unfeasibly large number of materials.
 3. Ensure that the energy above hull is less than 0.2 eV per atom.

For **unsuitable** candidates, we implement the same strategy as for the first Ferrenti approach. The result is a somewhat unbalanced set of labeled materials, with up to 75 % of the materials found in the suitable group. However, the labeled data encompasses 78 % more materials than the Ferrenti approach, consisting of 2141 suitable and 684 unsuitable materials.

The empirical approach

The third approach differs substantially from the others in terms of material labeling, as we apply knowledge from the field (see for instance Refs. [11, 13, 14, 40] for an overview) to pick our promising material host candidates. In other words, we restrict ourselves to materials where quantum compatible properties have been either experimentally demonstrated or theoretically predicted. We therefore name this scheme *the empirical approach*. Due to the concern of ending up with too small datasets for training and testing of the ML methods, we will also include some materials that are promising in the sense that they have suitable properties for accommodating deep level defects that can potentially exhibit quantum effects — even though such effects may not yet have been shown. These material choices were motivated by the criteria and predictions for QT compatible deep level defects defined by Weber *et al.* [8]. We have also included materials where properties like single-photon emission and spin manipulation have been observed but attributed to excitonic effects or quantum dots (formed by self-assembly or lithographic structuring) rather than being defect related.

Table I contains an overview of known semiconductor materials with demonstrated quantum compatible characteristics. The table forms the basis for picking suitable candidates for the empirical approach. The properties being studied arise from mechanisms related to, e.g., point defects, bound excitons, and both self-assembled and lithographically structured quantum dots and nanostructures such as 2D materials. Quantum emission signatures have been assigned to specific defects in both diamond and SiC, but for most other materials, secure identification of the responsible defects or structure related mechanism is still lacking.

The strategy for picking **suitable** candidates in the empirical approach can be summed up as;

1. Select candidate materials that match the formulas in Table I, or the formulas ZnSe, AlP, GaP, AlAs, ZnTe, CdS [8] and SiGe [41], as these materials have been predicted to behave as suitable quantum hosts based on favorable properties such as a wide band gap and low spin-orbit coupling.
2. Ensure that the candidates are present in the ICSD.
3. Perform a manual screening for appropriate crystallographic structures.

TABLE I: Overview of materials and defects that have been demonstrated to exhibit quantum compatible characteristics such as single-photon emission and coherent spin manipulation. The subscript denotes lattice site and V refers to a vacancy.

Material	Band gap (eV)	Defect candidates	References
Diamond	5.5	$\text{N}_\text{C}V_\text{C}$, $\text{Si}_\text{C}V_\text{C}$, $\text{Ge}_\text{C}V_\text{C}$	[42–47]
SiC	2.2–3.3	V_Si , $V_\text{Si}V_\text{C}$, $\text{C}_\text{Si}V_\text{C}$, $\text{N}_\text{C}V_\text{Si}$	[48–51] [8, 11, 52]
Si	1.1	P, G, unidentified	[53–55]
(2D) h -BN	6.0	Unidentified defects	[56–58]
(2D) MoS_2 , WSe_2 , WS_2	< 2.5 eV	Bound excitons	[40]
ZnO	3.4	Unidentified defects	[59]
ZnS	3.6 (zincblende)	Unidentified defects	[60]
GaAs	1.4	Quantum dots	[61]
GaN	3.4	Quantum dots, unidentified defects	[62, 63]
AlN	6.0	Unidentified defects	[64]

After the first stage of picking candidates we are left with a list of 202 matching formulas which includes 12 entries that have a band gap of less than 0.4 eV. These 12 entries are calculated to be thermodynamically unstable in terms of the energy above hull, and will decompose into other materials in the list — incidentally, the resulting structures all have calculated band gaps that are substantially larger than 0.5 eV. We include all of the 202 structures apart from C (mp-568410) which is a metal according to AFLOW-ML.

Entries matching the formulas C, SiC, BN, MoS₂, WSe₂ and WS₂ (quantum compatible characteristics have been demonstrated) were manually screened to see whether they have a matching structure to the respective candidates discussed earlier and summarized in Table I. For carbon, we admit three-dimensional diamond-like structures as explicitly stated in the column tags from the Materials Project. Additionally, we find several two-dimensional structures of carbon with a large band gap (> 1.5 eV) among the data. We add these as suitable candidates. Complex structures (e.g., C₂₈, C₄₈ and C₆₀) were moved to the test set in our machine learning studies. For SiC we admitted all entries which included the 2H, 3C, 4H, 6H and 15R polytypes. Concerning BN, MoS₂, WSe₂ and WS₂, we only admit two-dimensional structures. For non-matching structures not mentioned so far, we move them to the test set to see if they will be predicted as suitable or not by the ML methods that are applied in a later stage.

The materials AlP, GaP, AlAs, ZnTe and CdS were manually screened for tetrahedrally coordinated structures, and have been included since Weber *et al.* [8] identified them as potentially promising candidates due to suitable material properties. We note that only tetrahedrally coordinated structures of the given formulas were labeled as suitable after imposing a band gap restriction of 0.5 eV.

Following the three screening steps in the empirical approach, a total of 187 entries were labeled as suitable candidates for the training set. Notably, the candidates labeled as suitable in the empirical approach contain only compounds that are either elementary (unary) or binary. Since the set of labeled data constituting suitable candidates in the empirical approach contains rather few entries, we add 400 candidates that are labeled as **unsuitable**. These are picked at random from the pool of unsuitable candidates from the two previous approaches, in addition to those that were marked as unsuitable during the manual screening process.

Comparing the approaches

The three different approaches to data mining place a particular emphasis on each of their specific goals. The Ferrenti approach relies on choosing only elements with spin zero isotopes and compounds that crystallize in nonpolar spacegroups together with practical filters, while the extended Ferrenti approach allows a larger variety of elements and removes the practical reasons for excluding candidate materials. Thus, the first approach targets a more narrow prediction space than the second approach does. However, the most selective approach is the empirical one, relying exclusively on previous findings rather than assuming which material properties are influential.

Considering the materials present in the labeled datasets of the first two approaches more closely, we find that carbon in a diamond-like structure is classified as a suitable candidate in the Ferrenti approach, in good agreement with experimental observations (see Table I). Interestingly, in the labeled data from the Ferrenti approach, we find that carbon in two-dimensional graphite-like structures are marked as suitable as well. All structures of silicon are also labeled as suitable together with one entry of silicon carbide. Note that this is the 3C polytype of SiC, meaning that the most well-established quantum compatible SiC polytype, 4H-SiC, is missing. Among other potentially suitable candidates we find ZnS, ZnSe, ZnO and ZnTe present in the labeled data from the Ferrenti approach. In the labeled dataset of the extended Ferrenti approach, we find a single entry for each of SiC, Si, GaN, ZnS, GaP, AlAs and AlP, carbon in both diamond- and graphite-like structures, and AlN in three different configurations. The labeled dataset includes a larger variety of materials that are known to be quantum compatible as compared to the Ferrenti approach due to admitting more elements in the initial selection process. However, since we also included a more stringent band gap restriction of 1.5 eV (to reduce the amount of materials in the training set), we find a more sparse representation of each known chemical formula present in the labeled data.

The labeled data for each approach is visualized in Fig. 3 as parallel coordinate plots for a few selected features informed by the criteria proposed by Weber *et al.* [8]. Parallel coordinate schemes [65, 66] represent a multi-dimensional data tuple as one polyline crossing a parallel axis. The selected features are found on the x -axis, while the y -axis shows the value of the data. Thus, parallel coordinate plots can turn complex many-dimensional data into a compact two-dimensional representation. Due to possible data cluttering, the figure visualizes a random sample of each class (suitable or unsuitable) with an upper limit of 250 per class with transparent lines. The green and red polylines represent suitable and unsuitable candidates, respectively.

The Ferrenti approach and the extended Ferrenti approach share several similarities. For example, the entries in Fig. 3 span roughly the same dimensions for the Ferrenti and extended Ferrenti approaches, despite the more restricted nature of the former. Moreover, the two approaches' upper limits for ionic character and covalent range are very similar in the case of candidates labeled as both suitable and unsuitable. The most striking difference is found

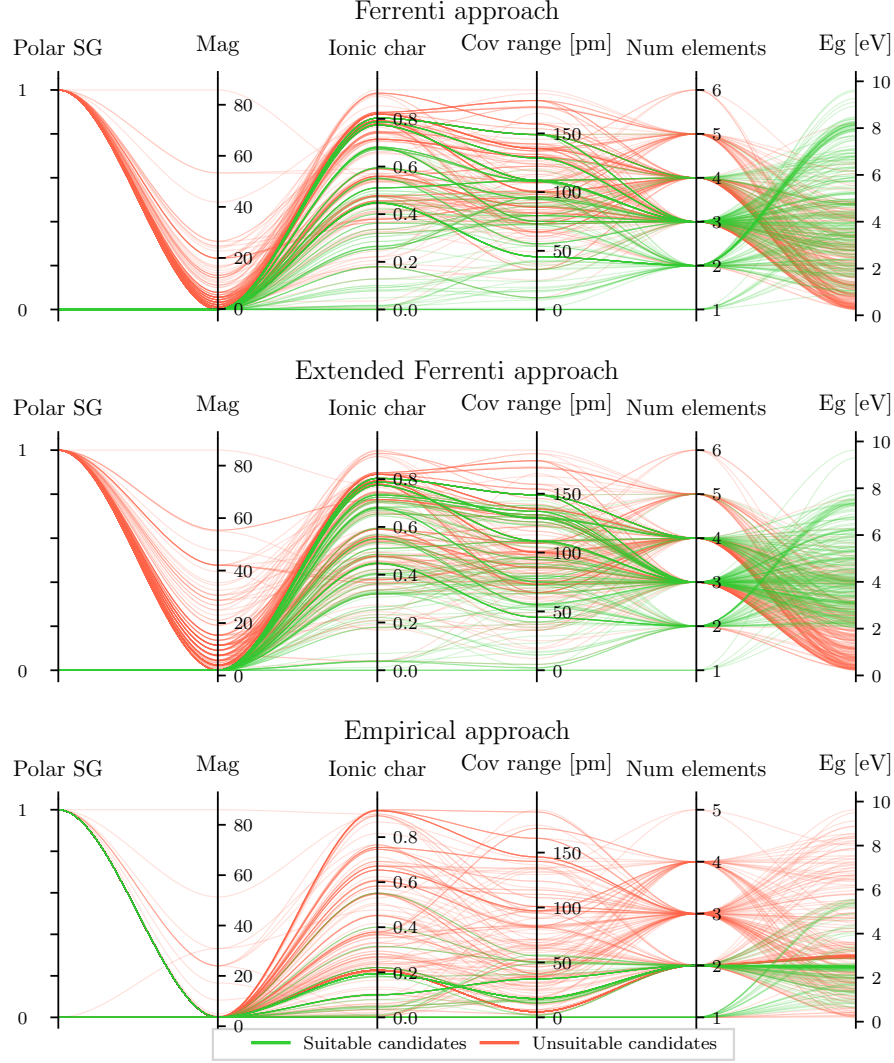


FIG. 3: Parallel coordinate plots for the different approaches. To limit data cluttering, we have randomly collected up to 250 entries for each class. For the empirical approach, we have used all the 187 suitable candidates. The axes show total magnetization (mag), space group (SG), ionic character (ionic char), covalent range (cov range) as calculated from elemental properties, number of elements (num elements) and energy gap (Eg) as extracted from the MP database.

for the empirical approach. Again, the entries labeled as suitable candidates do not exhibit any magnetization, but now the materials crystallize in both polar and non-polar space groups — contrarily to the two Ferrenti approaches. Moreover, we observe that the range of covalent radius and maximum ionic character span a substantially smaller parameter space in the case of the empirical approach than for the two former ones. Here, we can begin to deduce some trends, if the findings from the empirical approach are taken as guidance. In the empirical approach the polarity of the material’s space group may be less important compared to the Ferrenti based approaches, while the accepted values for ionic character and covalent range appear more restricted (Fig. 3).

To illustrate the composition of the labeled data we perform a principal component analysis (PCA) [67]. Here we apply PCA to the labeled datasets in order to reveal specific features that stand out. In its standard form, PCA relates the variance of the features with the eigenvalues of the covariance matrix [29, 30, 67]. We identify the two largest eigenvalues of the covariance matrix [29] of the complete initial dataset from the MP database, and transform the three labeled datasets according to the corresponding two eigenvectors. In essence, we reduce the dimensionality of the labeled data in order to visualize them. The result of this procedure is displayed in the scatter plots of Fig. 4. Note that some minor differences between the three panels in Fig. 4 may occur due to the process of removing the mean and scaling to unit variance. Red triangles represent unsuitable candidates while colored symbols (green, blue,

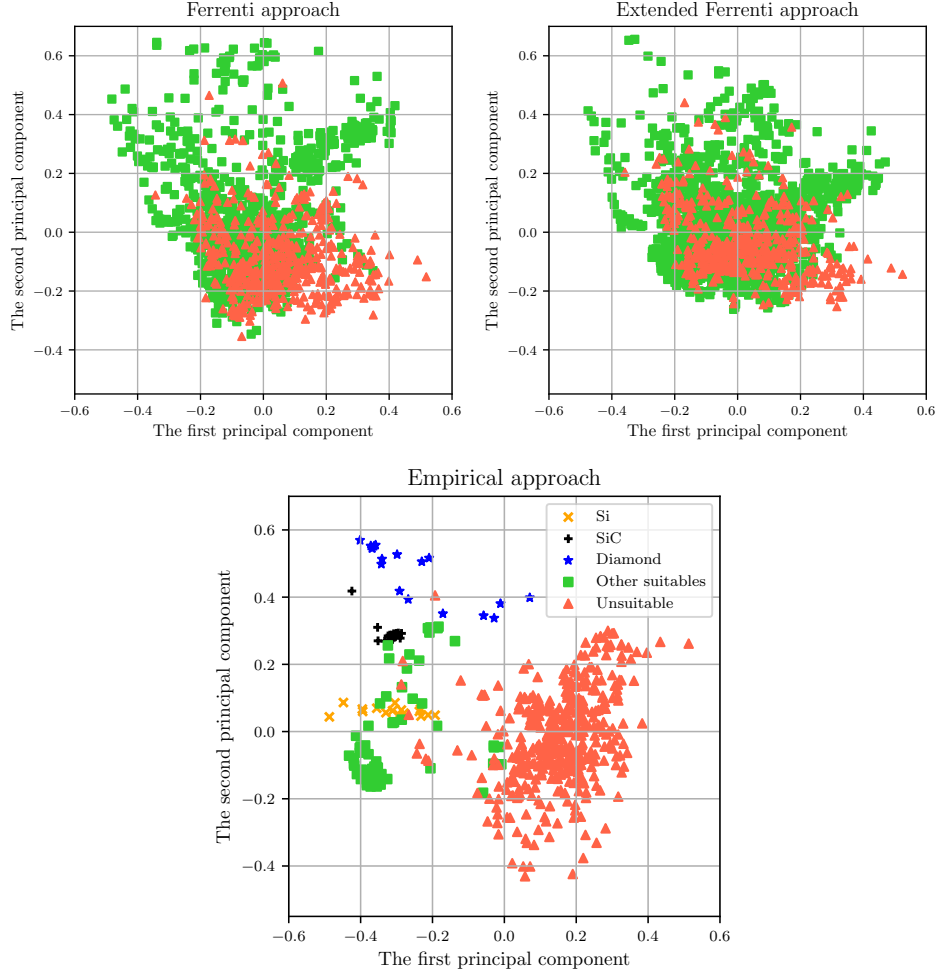


FIG. 4: Two-dimensional scatter plots for the three different approaches. We have identified the eigenvectors corresponding to the two largest eigenvalues of the covariance-matrix, that is, the two most important principal components of the initial data from the Materials Project query. Then, we have transformed the sets of labeled data resulting from the three approaches and visualized them as scatter plots. Green squares display suitable candidates, along with the black (SiC), blue (diamond) and yellow (Si) symbols in the bottom panel, and red triangles represent unsuitable candidates.

black and yellow) represent suitable candidates. Due to the complexity of reducing the large amount of features down to only two, both suitable and unsuitable candidates for the Ferrenti and the extended Ferrenti approaches are largely overlapping. The logic behind categorizing materials in two classes (suitable and unsuitable) appears to not translate into a distinct separation, at least not in the representation of Fig. 4 for the Ferrenti and extended Ferrenti approaches. Hence, using these approaches for predicting QT material hosts could prove challenging for any model that would try to glean a clear-cut boundary between materials that are and are not suitable for QT. We therefore expect that the two Ferrenti approaches could need supplementary dimensions for further distinguishing between the materials in the two categories (suitable and unsuitable), or that it may prove challenging to find a generalized model based on the training sets that were defined according to the above discussion.

In the case of the empirical approach, on the other hand, we can discern a trend where the upper left part of Fig. 4 is dominated by suitable candidates (green, blue, yellow and black symbols), while the unsuitable candidates (red triangles) are similarly restricted to the lower right corner, albeit with some exceptions. Interestingly, we observe that different configurations of the famously quantum compatible materials of diamond (blue), silicon carbide (black) and silicon (yellow) are grouped together but each material class is separated in its own region. This indicates that the empirical approach may be capable of separating materials based on their underlying properties. Thus, devising a logical framework for the data mining process is likely important in order to obtain suitable datasets for training and testing of the ML methods.

Machine learning and principal component analysis

We have applied four well-tested ML algorithms in order to classify specific materials as candidate systems for QT. These are logistic regression, decision trees, and the ensemble methods random forests and gradient boosting [28–30]. These four ML methods were applied to the labeled data sets that were optimized using each of the three approaches (Ferrenti, extended Ferrenti and empirical) outlined above, yielding twelve sets of candidates that were predicted as suitable material hosts for QT applications.

In order to reduce the dimensionality while preserving most of the variance in the data, we have performed a principal component analysis [67] for each of the three approaches (Ferrenti, extended Ferrenti and empirical). First, the eigenvalues of the covariance matrices of the three training sets are identified in order of decreasing value, resulting in three distinct PCA models (as opposed to the case of Fig. 4 where all panels are based on the same model). Next, in the evaluation of the ML methods for the three different approaches we apply a 5×5 stratified cross-validation [29] when searching for the optimal hyperparameter combinations. We note that all four ML methods have high evaluation metrics for the optimal hyperparameters. Further details on the evaluation of the ML methods and the results from the principal component analyses for all four methods are shown in the Supplementary Information [39]. The three data labeling approaches (Ferrenti, extended Ferrenti and empirical) define sets of labeled data of varying sizes, where the smallest is from the empirical approach. For small datasets, it is proven to be beneficial to repeat the cross-validation analysis [29], as this is a method that allows us to measure the stability of the predictions against perturbations (i.e., few different entries) in the training data [68].

Figure 5 visualizes different parameters for the most important principal components ranked in descending order by (a) the explained variance for the Ferrenti (upper panel) and empirical (lower panel) approach, and (b) the gradient boosting coefficients for the corresponding approaches. This differs from Fig. 4, where we visualize the two most important eigenvectors for all approaches that originates from the same covariance matrix. Note also that in Fig. 4 we performed a PCA analysis on all data. Here, the ML analysis is done on the labeled data for each approach. For the Ferrenti approach (Fig. 5a) to reach the 95 % accumulated explained variance, a total of 144 principal components must be included. For the extended Ferrenti approach, we find that gradient boosting performs optimally at 93 principal components. In the case of the empirical approach and in contrast to the Ferrenti approaches, decision trees and random forests exhibit the best performance for just a few principal components, and experience a considerable degree of overfitting when involving more principal components. Gradient boosting also experiences the best performance for just a few principal components.

Figure 5(b) reveals that the most important feature for the gradient boosting method in the Ferrenti approach is the fifth principal component (fifth largest eigenvalue). The trend is maintained for all four ML methods (see the Supplementary Information at [39]). Moreover, by selecting the highest values in this eigenvector, we find that the corresponding features originate from the DFT computed band gap of the elemental solids among the elements in the compound as calculated by the Materials Agnostic Platform for Informatics and Exploration (MagPie) [70]. The second most important principal component for the Ferrenti approach exhibits significant contributions from the covalent radius, the ionic character and the packing efficiency among the elements in the composition. The data originates from elemental calculations from MagPie and are aggregated as either minimum, mean, standard deviation, or maximum. While the first principal component encompasses the largest explained variance, it does not provide a clear and specific information on which features it represents.

The interpretation of feature importance for the extended Ferrenti approach is substantially more challenging than in the Ferrenti approach. We find for logistic regression and decision trees that no feature is different than any other in the cross-validation, mainly due to a large variety of accuracy. However, we find that random forests and gradient boosting both mark the fifth principal component as important. Similar to the Ferrenti approach, the corresponding features with the highest value for the fifth principal component originate from the DFT computed band gap of the elemental solids among the elements in the composition.

The empirical approach differs in many aspects from the Ferrenti and extended Ferrenti approaches. Firstly, we find that the number of principal components necessary to obtain 95 % variance is reduced to 103 components, which is 41 and 56 less than the Ferrenti and extended Ferrenti approaches, respectively. Thus, the variance of the training set is found to be described with fewer principal components, indicating a simpler model. Secondly, the first principal component is by far the most important feature for all ML methods in the empirical approach, as visualized in the lower panel of Fig. 5(b) for gradient boosting. Similar conclusions are reached when using logistic regression, decisions trees and random forests as classification methods (see the Supplementary Information at [39]). The distinct importance of the first principal component partly explains why we experience a high accuracy for only a single feature. The first principal component’s corresponding features is a complex combination of several material properties, but we find that it includes bond orientational parameters, coordination numbers, and the radial distribution function of a compound’s crystal system. The standard deviation of the radial distribution function appears multiple times in the list of features and is of particular importance. Thirdly, the empirical approach differs in how much explained

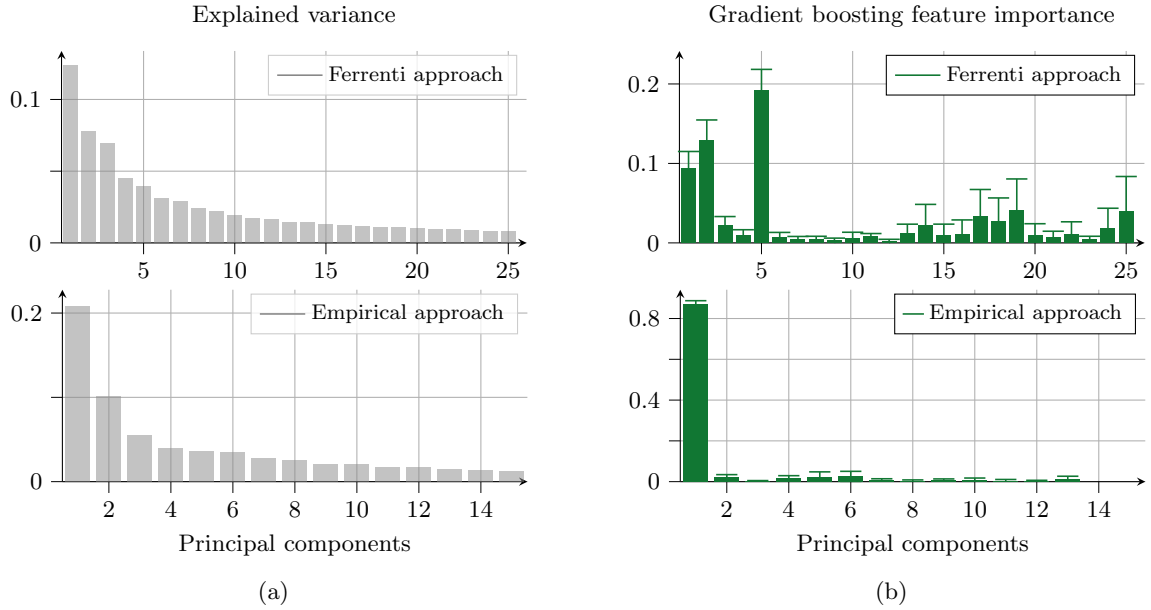


FIG. 5: Visualization of different parameters for the 25 most principal components ranked in descending order by (a) the explained variance for the Ferrenti (upper panel) and empirical (lower panel) approaches, and (b) the gradient boosting coefficients for the corresponding approaches. Note that in (a) we visualize only the results of the PCA analysis while in (b) there are results of the PCA reduced training sets using gradient boosting [29, 69]. The latter shows that for the empirical approach most of the physics is represented by a few features. The results in (b) are similar to those obtained with logistic regression, decision trees and random forests.

variance is retained by the first component, which is 21 %, while it is 14 % for the Ferrenti approach and 11 % for the extended Ferrenti approach. We find the difference striking considering that the approaches share the same ultimate goal, but where the training sets apparently exhibit large and significant variations.

Intriguingly, the principal components that are deemed important by the three approaches differ substantially. The Ferrenti and extended Ferrenti approaches place particular value on the material’s band gap and the ionic or covalent character of the bonding. Indeed, precisely these features were used as guidance in the data labeling process. Thus, the ML methods are preserving the criteria imposed in the Ferrenti approaches and perpetuate our initial assumptions on which material properties to look for. For the empirical approach, on the other hand, we did not guide the selection process in terms of specific properties. Here, the ML methods appear to be informed by other characteristics than band gap and ionic character that are more related to symmetry and crystal structure, based on the repeated appearance of bond length, orientation and radial distribution in the first principal component. In the empirical approach, the ML methods appear to be inferring that the common trend among the materials that are known to be suitable for QT is related to more complex mechanisms in the crystal structure and bonding.

Predicting Suitable Material Hosts for Quantum Technology

After training the ML algorithms using parts of the data labeled during the three data mining approaches, the ML methods were employed on the test sets. The number of candidates predicted by each of the four ML methods logistic regression, decision trees, random forests and gradient boosting is visualized in the Supplementary Information at [39] for each of the three approaches (Ferrenti, extended Ferrenti and empirical). An overview of the number of predicted suitable candidates as a function of a given ML algorithm’s confidence is summarized in Table II.

The Ferrenti approach

From the predicted dataset of 23.623 materials the four ML methods consistently predict at least 11.000 materials as promising candidates in the case of the Ferrenti approach. All four ML methods agree on a total of 6804 suitable candidates, however, many of these materials are predicted with an incidence similar to that of a coin flip. If we

TABLE II: Overview of the number of predicted suitable candidates as a function of a given ML method’s confidence when the four methods in an approach agree. A threshold value of 0.5 represents the confidence of a coin flip while 1.0 is a fully confident prediction.

	0.50	0.75	0.85
Ferrenti approach	6804	1784	258
Extended Ferrenti approach	9227	4735	2001
Empirical approach	214	66	9
All approaches together	47	6	0

were to raise the minimum confidence cut-off for a prediction to, e.g., 0.75 instead of 0.5, the above ML algorithms would only agree on 1784 suitable candidates. We find that all four ML methods admit almost all materials with a chemical formula matching the known suitable candidates (see Table I) that were present in the labeled data. This can allow materials with unfortunate structures to be labeled as suitable candidates. Notably, the ML methods do not maintain the band gap restriction from the training set definition, where all materials with band gaps lower than 0.5 eV were eliminated. This trend is not carried over to the predicted data. Indeed, we find that many entries with band gaps lower than 0.5 eV are predicted as suitable candidates by all four ML methods employed herein — despite the principal component analysis revealing that the band gap is an important feature for the machine learning classification in the Ferrenti approach. On the other hand, due to the known underestimation of band gaps by the PBE functional, the band gaps could in reality be larger for many of the materials.

The extended Ferrenti approach

Next, we turn to the more liberal extended Ferrenti approach containing a 78 % larger amount of labeled data than the Ferrenti approach. The four ML methods predict at least 13.000 materials as suitable candidates out of 22.550 materials in the unlabeled dataset, where they all agree on a total of 9227 predicted suitable candidates. Comparing to the labeled data reveals that all of the unlabeled candidates that are known to be suitable (see Table I) are, in fact, predicted as suitable candidates. However, the ML methods also predict materials as suitable that we would not expect according to, e.g., Weber *et al.* [8]. All four ML algorithms predict NaCl as a suitable candidate to confidences of 0.83 and 0.60 for two different configurations, despite the strong electrostatic interactions between Na and Cl and the ionic character of their bonding. Furthermore, although we enforced a conservative band gap restriction of 1.5 eV for the data labeling in the extended Ferrenti approach, we still find that all four implemented ML methods predict suitable candidates that exhibit band gaps substantially lower than 0.5 eV. As seen above, the ML methods applied to data labeled in the Ferrenti and extended Ferrenti approach are recognizing the band gap and bonding character as important, but the resulting predicted materials do not strictly follow the anticipated guidelines.

The empirical approach

The ML methods that were trained on the data extracted in the empirical approach predict substantially fewer candidates as compared to the other two approaches. A total of 842, 1197, 543 and 596 materials are classified as suitable candidates by logistic regression, decision trees, random forests and gradient boosting, respectively. All the four ML methods agree on a total of 214 suitable candidates to 0.5 confidence. Note that 51 of these have a band gap of 0.5 eV or smaller as reported in Materials Project [17, 18] from PBE-level DFT calculations. Increasing the threshold to 0.75 or 0.85 yields 66 or 9 predicted suitable candidate materials in the empirical approach, respectively (see Table II).

Consider the 9 materials that were classified as suitable to a confidence of 0.85 or higher by all four ML methods in the empirical approach; BN, CdSe (2 structures), BC₂N (2 structures), InAs, CuI (2 structures), and ZnCd₃Se₄. The nine materials (considering different crystal structures) each belong to one of the four crystal systems cubic, hexagonal, tetragonal and orthorhombic. Figure 6 visualizes the four different crystal systems while Table III lists important material properties of the relevant materials as reported in the MP database. Interestingly, we find that all nine materials appear to be four-fold coordinated, and that the first two lattice vectors (*a* and *b*) are identical in all nine cases whereas *c* may differ. There appears to be a high degree of symmetry present for all materials, but we find no elemental and perfectly symmetric semiconductors in the list.

Turning to each of the individual material predictions, the compound BN (mp-1639) was in fact present in the training data as a suitable candidate, and we believe this to be a partial explanation for why the ML methods

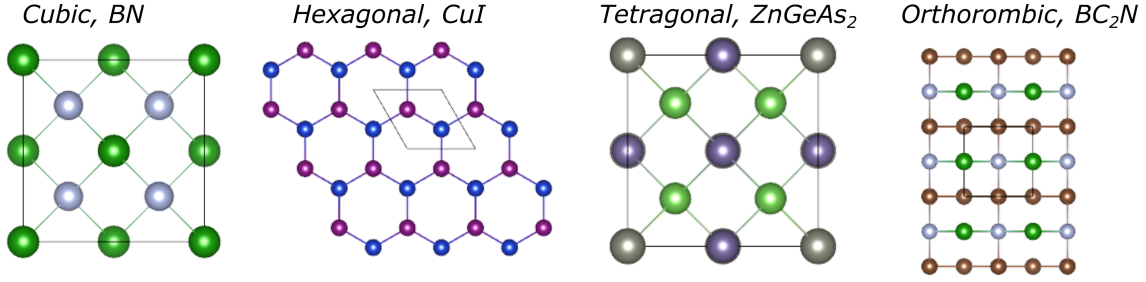


FIG. 6: Illustrations of the crystal systems that the 9 materials predicted by the empirical approach to 0.85 confidence, and the 6 materials predicted by all approaches to 0.75 confidence, may belong to. Four different symmetry categories are observed; cubic, hexagonal, tetragonal and orthorhombic. The viewpoints for all materials are down along the c -axis.

recognize BN as a suitable candidate with high probability. Furthermore, two compositions of CdSe (the cubic mp-2691 and the hexagonal mp-1070) have been predicted as suitable, possibly as a consequence of the similar compound CdS being labeled as a suitable candidate in the training set. The two compounds of CdSe share similar properties with calculated band gaps in the MP database of 0.5 eV and 0.6 eV, respectively.

In the case of BC_2N we find two compositions with the same chemical formula; the orthorhombically coordinated (mp-629458) and the tetragonally structured (mp-1008523) BC_2N . The former takes on a polar space group while the latter does not. The band gaps are listed as 1.9 eV and 1.7 eV, respectively, in the MP database. Both structures exhibit, as expected, a strong covalent character and have been studied for applications as nanostructures for electronic devices [71], hydrogen storage [72] and superhard materials [73, 74]. Importantly, the diamond-like structure of BC_3N was recently predicted as a promising spin qubit material host [75]. By creating a boron vacancy one can in theory obtain a defect center with similar properties to those found for the NV center in diamond. Whether this is also possible for BC_2N remains to be seen. We note that BC_3N was not represented in the MP database at the time of data extraction and is therefore not included in our dataset.

The compounds InAs (cubic, mp-20305), CuI (cubic, mp-22895 and hexagonal, mp-569346) and ZnCd_3Se_4 (cubic, mp-1078597) are listed in the MP database with band gaps of 0.3, 1.2, 1.2 and 1.7 eV, respectively. To the best of our knowledge, ZnCd_3Se_4 has yet to be synthesized, while self-assembled InAs quantum dots are exciting possible materials to use in quantum technology [76]. CuI has recently been synthesized as single-crystal epitaxial films and was shown to exhibit remarkable optoelectronic properties [77]. Interestingly, the material exhibits a large ionic character with more similarities to some of the oxide compounds in the dataset than, e.g., the materials that are more covalent in character like Si, SiC and diamond. The prediction of CuI in two configurations indicates that ionic character alone is not an obstacle for a material to be quantum compatible. It is unknown at this time whether any potentially favorable properties of CuI would originate from deep level defects or nanostructuring (e.g., quantum dots). NaCl was not predicted as suitable for QT in the empirical approach.

Out of the nine predicted materials (threshold of 0.85) we note that both instances of CuI are stated in the MP database to decompose into trigonal CuI, which was not present in our predictions. Hexagonal CdSe decomposes into cubic CdSe (in the list of 9), the cubic BN decomposes to hexagonal BN (mp-984, not in the predictions), and ZnCd_3Se_4 decomposes to $\text{ZnSe} + \text{CdSe}$ (in the list). Finally, both compounds of the BC_2N structure are listed to decompose into hexagonal BN + C. However, synthesis may still be possible depending on the growth conditions.

Lowering the cut-off requirement from 0.85 to 0.75 for all ML methods results in 66 candidate materials for the empirical approach. The full list of these 66 candidates is displayed in the Supplementary Information at [39]. In addition to the nine materials discussed above and some elemental and binary semiconductors, the list of 66 predicted suitable candidates now also includes ternary compounds of the formula ABC_2 . For the ABC_2 structures, the elements Ga, Cd, In and Zn can occupy the A-site, Cu, Sn, Ag and Ge take the B-site, while S, Se, Te, P or As may reside at the C site. Most of the predicted compounds include at least one toxic element with one exception: ZnGeP_2 (mp-4524) in a chalcopyrite-like tetragonal crystal structure with an indirect band gap of 1.2 eV [78] as reported in the MP database. In comparison, the experimentally reported band gap is somewhat larger at 2.0 eV [79]. ZnGeP_2 crystallizes in a non-polar space group, possesses no magnetic moment, exhibits strongly covalent bonds, and has been reported as an excellent mid-IR transparent crystal material that is suitable for nonlinear optical applications [78]. Importantly, it is possible to integrate sources of photon quantum states based on nonlinear optics with ZnGeP_2 [80]. We therefore identify ZnGeP_2 as an eligible candidate material for QT, but it remains to be seen whether the candidate can facilitate, e.g., the isolated deep energy levels often associated with defects exhibiting quantum compatible properties, or instead be a candidate for nanostructure based QT.

TABLE III: Overview of material properties as listed in the MP database for the 9 materials predicted by the empirical approach (> 0.85) and the 6 materials predicted by all approaches (> 0.75). Compounds marked by an asterisk are predicted to be unstable to transformation into other phases in the MP database.

	Material	Crystal structure	MP code	Density (g/cm ³)	Band gap (eV)	a, b, c (Å)	α, β, γ (°)
Empirical approach to 0.85 confidence	CdSe	Hexagonal	mp-1070	5.3	0.6	4.4, 4.4, 7.2	90, 90, 120
	CuI	Hexagonal	mp-569346	5.8	1.2	4.3, 4.3, 7.0	90, 90, 120
	CuI	Cubic	mp-22895	5.8	1.2	4.3, 4.3, 4.3	60, 60, 60
	CdSe	Cubic	mp-2691	5.3	0.5	4.4, 4.4, 4.4	60, 60, 60
	BN	Cubic	mp-1639	3.5	4.6	2.6, 2.6, 2.6	60, 60, 60
	InAs	Cubic	mp-20305	5.3	0.3	4.4, 4.4, 4.4	60, 60, 60
	ZnCd ₃ Se ₄	Cubic	mp-1078597	5.3	1.7	6.1, 6.1, 6.1	90, 90, 90
	BC ₂ N	Tetragonal	mp-1008523	3.3	1.6	2.6, 2.6, 3.7	90, 90, 90
	BC ₂ N	Orthorhombic	mp-629458	3.4	1.8	2.5, 2.6, 3.6	90, 90, 90
All approaches to 0.75 confidence	CdSnP ₂	Tetragonal	mp-5213	4.6	0.7	7.2, 7.2, 7.2	131.1, 131.1, 71.7
	GeSe	Cubic	mp-10759	5.5	0.4	4.0, 4.0, 4.0	60, 60, 60
	InP	Cubic	mp-20351	4.6	0.5	4.2, 4.2, 4.2	60, 60, 60
	InP	Hexagonal	mp-966800	4.6	0.5	4.2, 4.2, 6.9	90, 90, 120
	SiSn	Cubic	mp-1009813	4.4	0.4	4.3, 4.3, 4.3	60, 60, 60
	ZnGeAs ₂	Tetragonal	mp-4008	5.2	0.6	7.0, 7.0, 7.0	131.4, 131.4, 71.2

Among the 66 materials that all ML methods in the empirical approach agreed on to a 0.75 confidence level, we also emphasize the presence of interesting materials like Ge (in three configurations mp-137, mp-1067619 and mp-1198022), GeC, BP and InP. The element Ge in the cubic structure (mp-1198022) shares many properties with Si and C in addition to the periodic column number. Germanium has the highest hole mobility of all semiconductors at room temperature, and is therefore considered a key material for the process of extending the chip performance in classical computers beyond the limits imposed by miniaturization [81]. Furthermore, like SiC, device design based on Ge can take advantage of the mature large-scale fabrication of silicon due to the material's comparable properties. Similar considerations could be made for the case of GeC. Data from the MP database suggests that the cubic compound GeC (mp-1002164) is a covalently bonded semiconductor having a band gap of 1.8 eV. Consequently, with SiC being widely known as a highly suitable host material for QT compatible defects, we encourage further research on GeC due to its similarities with SiC. Next, the compound BP (mp-1479 and mp-1008559) is present in our predictions in both the cubic and hexagonal structures, with indirect band gaps calculated at 1.5 eV and 1.1 eV, respectively. Both configurations of BP are nonmagnetic, non-toxic and BP has been synthesized with a potential for large-scale production [82]. Lastly, InP (mp-966800) was also predicted as a suitable candidate to 0.75 confidence. The compound inhabits the hexagonal structure with corner-sharing InP₄ tetrahedra. InP is reported in the Materials Project to have a direct band gap of 0.5 eV and is considered as one of the most promising candidates to compete with Cd- or Pb- based QDs for, e.g., display and lighting applications [83, 84]. We underline the possibility of using InP-based quantum dots for QT applications.

We note that all ML methods in the empirical approach agree on several oxides being potential candidates when the prediction cut-off is set low enough (but still above 0.5). However, we find that almost all the oxide compounds fall between the decision boundary defining suitable and unsuitable candidates, and none are present in the list containing the 66 suitable candidates using a 0.75 threshold. Due to the labeling of ZnO as a suitable candidate in the training set, we believe that the boundary was shifted sufficiently to admit several oxides as suitable candidates. Removing ZnO from the training set for the empirical approach might result in fewer or even no oxide compounds being predicted above a suitability limit of 0.5 and could be an interesting future pathway for investigation.

Comparing to the work of Ferrenti *et al.* [27], they suggest a list of 541 viable hosts after the data mining procedure. Among these, we find only a single material present in our list of 66 candidates predicted by the four ML methods in the empirical approach: the nontoxic compound MgSe (mp-10760) which crystallizes in the rock-salt structure, is expected to have a 2.0 eV band gap and decomposes to a similar MgSe configuration. MgSe is notable for its available spin-zero isotopes in accordance with the criteria set by Weber *et al.* [8] and Ferrenti *et al.* [27]. We note that these properties may favor defects acting as spin centers with qubit potential and MgSe is thus identified as an interesting host material in this regard.

Comparing the approaches

Comparing the three approaches Ferrenti, extended Ferrenti and empirical (see also Table II), we observe that the extended Ferrenti approach is the least restricted also for the predicted materials and admits the largest number

of entries regardless of the confidence threshold. The Ferrenti approach also predicts a large amount of materials as suitable compared to the empirical approach, but the resulting data are otherwise not very different from the extended version. The empirical approach, on the other hand, predicts fewer suitable candidates that are possible to examine manually. Manual verification through a literature survey will often not be possible, however, and perfecting automatic data mining and analysis is therefore an important goal of material informatics [85].

Despite notable differences, as mentioned above, there is substantial overlap of predicted materials between the three approaches. We find that 119 of the 214 candidates predicted as suitable by the empirical approach (threshold at 0.5) were also selected by all ML methods in the extended Ferrenti approach. Similarly, 78 of the 214 materials are also predicted as suitable by all ML methods in the Ferrenti approach. All approaches and their corresponding ML methods agree on a total of 47 potential candidates to 0.5 confidence, where 8 are elementary (unary), 29 binary, and 10 tertiary (see the Supplementary Information at [39] for the full list). Several interesting materials that were also discussed above for the predictions by the empirical approach are present among these 47, including BP, CdSe, GeC, InP and Ge. However, we also find certain materials that cannot be classified as semiconductors, such as P, I, N₂ and H₂. We note that none of these were included by the empirical approach when the threshold was set to 0.75 or above.

Importantly, there are 6 materials that all approaches (Ferrenti, extended Ferrenti and empirical) and ML methods agree on above a 0.75 threshold. These are ZnGeAs₂ (tetragonal mp-4008), CdSnP₂ (tetragonal mp-5213), GeSe (cubic mp-10759), InP (cubic mp-20351), InP (hexagonal mp-966800) and SiSn (cubic mp-1009813). Here, we can distinguish three groupings in crystal structure and they agree with those discussed above; cubic, hexagonal and tetragonal (illustrated in Fig. 6). The relevant material properties are summarized in the lower part of Table III. Note that hexagonal InP is listed in the MP database as decomposing into cubic InP while SiSn is stated to decompose into Si + Sn. We can deduce a series of similar trends as for the 9 materials from the empirical approach — e.g., the same space groups are manifesting, we find only binary or tertiary compounds, and $a = b$ for all materials while c is different for hexagonal InP (unstable). However, consistent four-fold coordination and the bond angle trends we observed for the first 9 materials are not maintained as GeSe is six-fold coordinated. We highlight these six compounds, along with the nine predicted by the empirical approach to 0.85 confidence, as particularly interesting for future in-depth theoretical and experimental studies.

DISCUSSION

Taking a closer look at the reasoning behind the choices made by the different ML methods during the classification process, we can start to identify important driving forces for manifestation of quantum compatible properties in semiconductors. The analysis of the principal components extracted from the ML methods revealed that the most important principal component for both the Ferrenti and extended Ferrenti approaches encompasses features related to the band gap and chemical environment. This means that the band gap criterion imposed in the training set selection is at least somewhat satisfied. In other words, the Ferrenti approaches appear to reproduce the logic of the initial selection process. For the empirical approach, on the other hand, band gap related features were not recognized as important in the dominant principal component.

Figure 7(a) displays the number of predicted suitable materials as a function of band gap (from the MP database), and reveals that all approaches predicted a substantial amount of materials with a low band gap (below 0.5 eV). Moreover, the distribution of materials that were identified as suitable is rather broad with the Ferrenti approaches exhibiting a peak around a 2.5 eV band gap. Coincidentally, the band gap of, e.g., 4H-SiC is usually computed at around 2.5 eV using the PBE functional. The empirical approach exhibits a more scattered data distribution with local maxima around 1-2 and 4 eV.

The secondmost important principal component in the Ferrenti approach encompasses properties such as the ionic character, covalent radius and maximum packing efficiency (see the Supplementary Information at [39] for the predicted material distributions of the latter two features). Intriguingly, as shown in Figure 7(b), the predicted candidates distribute over a broad range of ionic characters for all three approaches - even peaking at a relatively high ionic character of 0.8. We note that this may be a result of the distribution of the initial dataset of 25,000 materials that peaks around a similar value (not shown). The minor peak in the empirical approach's predicted materials around 0.4 ionic character is not present for the Ferrenti approach, extended Ferrenti approach nor the overall data distribution. For comparison, all SiC entries in the dataset have maximum ionic characters of ~ 0.1 . Further, we find that the covalent radii of the materials (see the Supplementary Information) exhibit two distinct peaks in the data distribution. The trend of two data peaks is repeated for the maximum packing efficiency but is much more prominent for the empirical approach. This indicates that the material density, or in other words the bond length, is an important parameter for QT suitability.

The ML methods in the empirical approach consistently identified the first principal component as the predominant

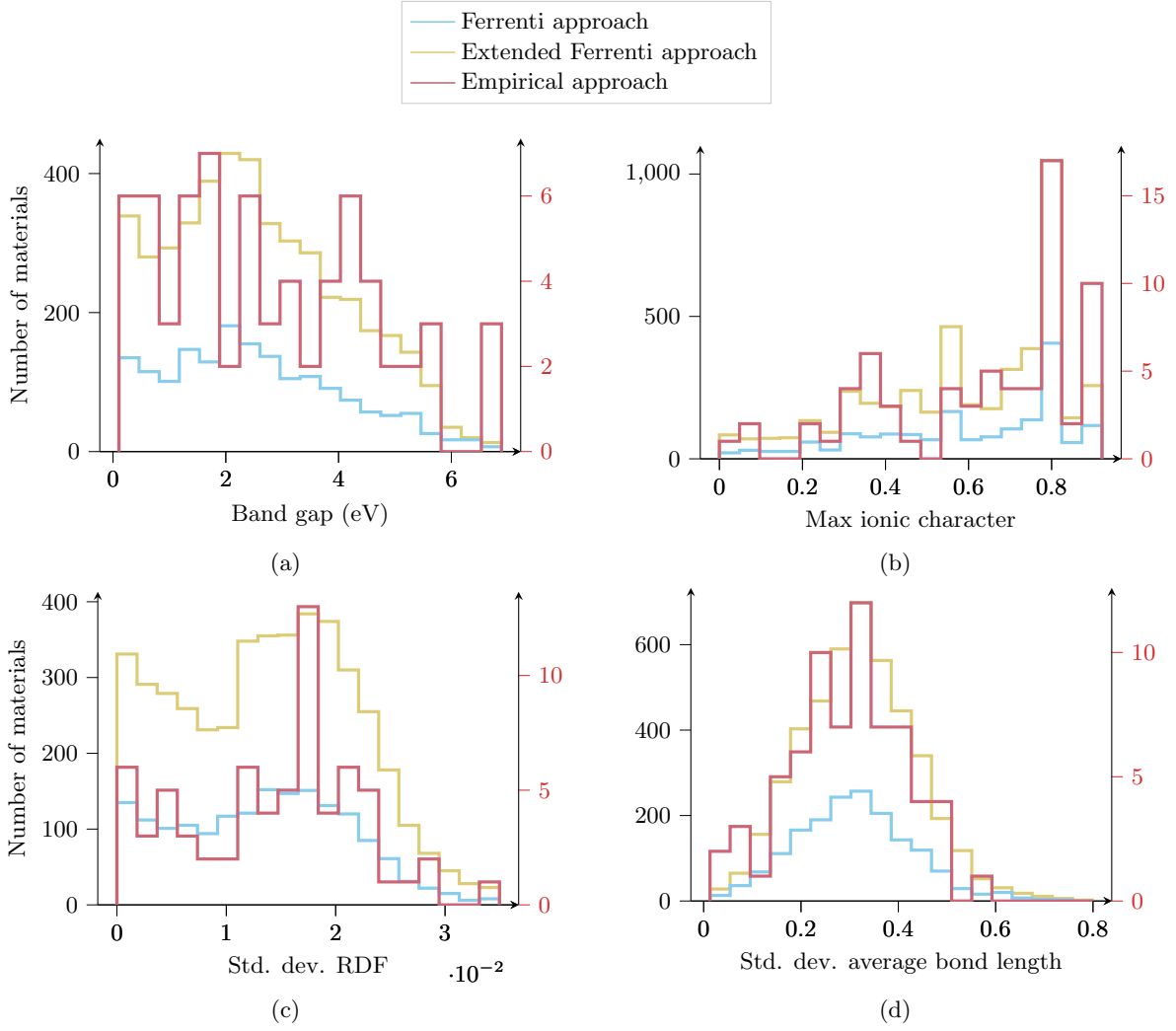


FIG. 7: Histograms of predicted suitable materials as a function of the (a) PBE-calculated band gap from Materials Project, (b) maximum ionic character, (c) standard deviation of the radial distribution function (RDF) and (d) standard deviation of the average bond length. The total number of predicted materials is 6804 for the Ferrenti approach, 9227 for the extended Ferrenti approach, and 214 for the empirical approach. The Ferrenti and extended Ferrenti approaches refer to the left y -axis and the empirical approach to the right.

one. Identifying the single most important feature in this principal component proved challenging as it is the combined impact of several features that is important. Here, we find that the standard deviation of the radial distribution function (RDF) has a particularly strong impact since it appears four times in different forms in the top ten list over dominating features. One configuration of the standard deviation of the RDF is demonstrated in Fig. 7(c), with two others being included in the Supplementary Information at [39]. Intriguingly, the standard deviation of the RDF exhibits substantial discrepancies between the Ferrenti approaches and the empirical approach. This reflects the observations from the principal component analysis where the first principal component is dominated by the RDF and similarly symmetry related features. For the empirical approach, there is a sharp peak in the preferred value for the standard deviation for the RDF, while the Ferrenti and extended Ferrenti approaches display an even distribution across a broader range.

We interpret the standard deviation of the RDF such that zero standard deviation in the RDF means that there is no variation in the radial symmetry throughout the material. Similarly, zero standard deviation in the average bond length would mean that all bonds are identical throughout the crystal. Intriguingly, the peaks in the distributions of the predicted materials are not found for perfectly symmetric materials with identical bond lengths; instead, some variation in the bond and wavefunction distributions is found to be optimal for a material to be suitable for QT. Note that the maxima in Fig. 7(c) seem to appear for moderate standard deviations in the RDF, indicating that a certain

degree of symmetry is necessary for a material to act as a suitable QT host. The exact degree and type of symmetry is still open to debate and merits further study. Similar symmetry related features such as the standard deviation of the average bond length (see Fig. 7d), the site fingerprint of the chemical environment and the bond orientation (see the Supplementary Information at [39]) are also influential in guiding the ML algorithms upon classifying suitable and unsuitable materials for QT in the empirical approach.

Interestingly, none of the nine materials predicted by the empirical approach above a 0.85 threshold, nor of the six materials predicted by all approaches to 0.75 confidence, are elemental. At most three different elements are present, but the emphasis is clearly on binary compounds. This resonates with the observations from the principal component analysis; an optimal degree of crystalline order likely exists for a material to manifest QT compatible properties, but some variations throughout the crystal in, e.g., bond length and symmetry are needed. This is in contrast to the expectations that guided the formulation of the test and training sets in the Ferrenti and extended Ferrenti approaches. Where most previous works have highlighted features such as band gap, polarity and ionic character as vital for a semiconductor to manifest quantum compatible features, our results reveal that local variations in the crystal structure related to symmetry and bond angles, length and orientation should be considered as well and may be more important.

Considering the trends in machine learning selectivity in light of the specific defect centers we know to be quantum compatible reveals fascinating characteristics. None of the known quantum emitters or spin centers seem to appear in completely uniform systems. For example, in the high symmetry materials of diamond and silicon, QT compatible defects are not intrinsic; neither the silicon vacancy in silicon nor the carbon vacancy in diamond, for example, have exhibited single-photon emission or controllable spin coherence. Instead, we find that quantum effects often appear after impurities are introduced, as evidenced by the phosphorous and carbon impurities in silicon [55, 86] and the nitrogen-vacancy, germanium-vacancy, tin-vacancy and lead-vacancy centers in diamond [87]. For the binary system of silicon carbide in different polytypes, on the other hand, intrinsic defects like the silicon vacancy and the divacancy are appropriate for our goals. While these trends have yet to be verified on a grander scale, our findings provide strong indications that local variations in the crystal structure are paramount for QT compatible properties to manifest in a semiconductor host.

Our findings are corroborated by the fact that all four ML methods predict a small set of the unlabeled materials as suitable, while agreeing on a large part of these materials. The methods we have chosen are all well tested, with random forests and gradient boosting methods tending to outperform the others, resulting normally in a small bias and variance [28–30].

To summarize, the data extraction tools developed in this work resulted in a dataset of over 25.000 materials and more than 4800 physics-informed features. Three different approaches (Ferrenti, extended Ferrenti and empirical) for data labeling were employed for training and testing of four ML algorithms. We find a much clearer separation between suitable and unsuitable candidates after data labeling in the empirical approach than for the two Ferrenti ones. Additionally, substantially fewer principal components are needed for obtaining optimal performance of the ML algorithms for the empirical approach. The principal component analyses of the ML methods’ performance imply that the Ferrenti and extended Ferrenti approaches, with their strong focus on band gap and bonding character when assessing a material’s quantum compatibility, largely replicate the specifications of the data labeling process. Valuable insight is gained from the empirical approach which highlights the importance of symmetry related properties in the bond orientations and wavefunctions.

The findings presented here firmly establish that material informatics (from data mining via featurization to machine learning predictions) is a viable and important route to new discoveries in important fields. Our focus has been on predicting new candidate materials to host single-photon emitters and spin centers for quantum technology applications, but the framework we have developed is suitable for other fields as well. We suggest two possible paths to further exploit our findings. One aspect is the pursuit of experimental verification of QT compatible effects in the materials predicted as suitable by the four ML methods. Another, perhaps even more important, route is to use the features and trends identified during the data mining and prediction process to understand the distinct material characteristics that enable quantum effects to manifest, opening thereby up for new discoveries in the field of quantum technologies.

METHODS

Databases

The Materials Project [17, 18] is an open-source project containing ground state properties of materials calculated using density functional theory (DFT) as implemented in the Vienna Ab initio Simulation Package (VASP) [32]. The Perdew-Burke-Ernzerhof [33] (PBE) functional is used to calculate band structures, while for transition metals, a

$+U$ correction is applied to correct for correlation effects [88]. The project is known as the initiator of materials genomics and offers a variety of calculated properties for over one hundred thousand inorganic crystalline materials, with frequent updates and extensions. Data extraction from Materials Project was performed in December of 2020 for the Ferrenti and extended Ferrenti approaches, and in March 2021 for the empirical approach. Therefore, the initial dataset for the two former approaches includes 77 more materials than that for the empirical approach due to erroneous entries that have been removed from the Materials Project database.

The Open Quantum Materials Database (OQMD) [19, 20] contains thermodynamic and structural properties of more than 600.000 materials. The calculations are performed with the VASP software and the electron exchange and correlation are described with the PBE functional. The $+U$ extension is included for several calculations considering specific elements [89]. Data extraction from OQMD was done in February of 2021.

JARVIS-DFT [21] is an open-source database based on the VASP software and consists of roughly 40.000 three-dimensional materials using the vdW-DF-OptB88 (OPT) functional [90, 91]. Structures included in the database are originally taken from the Materials Project [17, 18], and then re-optimized using the OPT functional. Finally, the combination of the OPT and modified Becke-Johnson (mBJ) functionals [92] is used to obtain a representative band gap for each structure [93]. Data extraction from JARVIS-DFT was done in January of 2021, where we utilized the version made available on 30.04.2021 (see Ref. [21]).

The AFLOW [22–24] repository is an automatic software framework for the calculations of a wide range of inorganic material properties. They utilize the PBE functional within VASP to relax and optimize the structure from the ICSD. Data extraction was performed in the period of January to February of 2021.

AFLOW-ML [25] is an application programming interface (API) that uses machine learning to predict thermo-mechanical and electronic properties based on the chemical composition and atomic structure alone, which are denoted as *fragment descriptors*. Initially, the API decides whether a given material is a metal or an insulator, where the latter is confirmed with an additional regression method to predict the band gap. The accuracy is validated by a five-fold cross-validation process for each ML method, where they report a 93 % prediction success of their initial binary classification method. In this work we utilized the Property Labeled Material Fragments (PLMF) openly available at their website (see Ref. [25]). We extract the crystallographic information files (CIF) for the crystals from Materials Project, use the CIF files as input to AFLOW-ML, which then returns an anticipated band gap. This process was executed during January of 2021.

Material Informatics

Matminer [26] is an open source toolkit for material analysis written in Python. Matminer provides modules to extract information from a wide variety of databases. Additionally, they provide the tools to construct possibly thousands of features from calculations based on a material’s composition, structure and electronic properties from DFT calculations, and have frameworks for visualization and automatic machine learning. To apply Matminer’s featurization tools, we extend an existing implementation by Breuck *et al.* [94], which was used to generate a supervised machine learning framework called the MODnet. The implementation by Breuck *et al.* provides featurization for a material’s composition, structure and atomic sites. However, Matminer also provides featurization tools for a material’s density of states (DOS) and band structure. Therefore, we extend their implementation to facilitate such featurizations. The features selected for featurization herein are summarized in the Supplementary Information at [39].

Pymatgen, a robust and open-source Python library for material analysis [35], was also employed to extract and generate features for several of the databases mentioned above.

Machine Learning

Machine learning represents the science of giving computers the ability to learn without being explicitly programmed. The idea is that generic algorithms exist which can be used to find patterns in a broad class of datasets without having to write code specifically for each problem. The algorithm builds its own logic based on the data.

The approaches to machine learning are many, but are often split into two main categories: supervised and unsupervised. In supervised learning we know the answer to a problem, and let the computer deduce the logic behind it. On the other hand, unsupervised learning is a method for identifying patterns and relationships in datasets without any prior knowledge of the system. Many researchers also operate with a third category, namely reinforcement learning. This is a paradigm of learning inspired by behavioral psychology, where learning is achieved by trial-and-error, solely from rewards and punishment. In this work our focus is on supervised learning only with labeled data for classification problems.

In this work we have applied four well-known and tested ML methods for classification problems, these are (see for example [28, 29] for discussions and applications):

1. Logistic regression,
2. Decision trees,
3. Random forests,
4. Gradient boosting.

Logistic regression [29] is a simple and frequently used method for binary and multi-category classification problems. In addition to logistic regression, we have also applied and tested the predictions made by decision trees and ensemble methods like random forests and gradient boosting, the latter through the application of the computationally efficient XGBoost library [69]. Gradient boosting and random forests use decision trees as weak learners and improve their predictability. For random forests this is implemented through a collection of randomized decision trees where a subset of the features in the datasets are selected randomly when building a decision tree. Boosting methods like gradient boosting use decision trees as weak learners and improve upon these by an iterative process that involves the estimation of the gradients of the cost/loss function [29]. Pure decision trees can easily lead to overfitting of the data under study, leading to a ML method that exhibits a high variance. Ensemble methods like random forests and gradient boosting on the other hand tend to soften the overfitting problem, resulting in both a small bias and a reduced variance of the employed method, see for example Refs. [28, 29] for an in-depth discussion of the bias-variance trade-off in machine learning. Gradient boosting implemented through the XGBoost library [69] is widely used by data scientists to achieve state-of-the-art results on many machine learning challenges.

DATA AVAILABILITY

The datasets that support the findings of this study are available online at [31].

CODE AVAILABILITY

The codes employed to develop the machine learning results are available online at [31].

-
- [1] D. Le Sage, L. Arai, D. R. Glenn, S. J. DeVience, L. M. Pham, L. Rhan-Lee, M. D. Lukin, A. Yacoby, A. Komeili, and R. L. Walsworth, Optical magnetic imaging of living cells, *Nature* **496**, 486 (2013).
 - [2] R. Ursin, F. Tiefenbacher, T. Schmitt-Manderbach, H. Weier, T. Scheidl, M. Lindenthal, B. Blauensteiner, T. Jennewein, J. Perdigues, P. Trojek, B. Ömer, M. Fürst, M. Meyenburg, J. Rarity, Z. Sodnik, C. Barbieri, H. Weinfurter, and A. Zeilinger, Entanglement-based quantum communication over 144 km, *Nature Physics* **3**, 481 (2007).
 - [3] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, and et al., Quantum supremacy using a programmable superconducting processor, *Nature* **574**, 505 (2019).
 - [4] J. Chow, O. Dial, and J. Gambetta, IBM quantum breaks the 100-qubit processor barrier (2021).
 - [5] H. J. Kimble, The quantum internet, *Nature* **453**, 1023 (2008).
 - [6] I. Aharonovich, D. Englund, and M. Toth, Solid-state single-photon emitters, *Nature Photonics* **10**, 631 (2016).
 - [7] D. D. Awschalom, R. Hanson, J. Wrachtrup, and B. B. Zhou, Quantum technologies with optically interfaced solid-state spins, *Nature Photonics* **12**, 516 (2018).
 - [8] J. R. Weber, W. F. Koehl, J. B. Varley, A. Janotti, B. B. Buckley, C. G. Van de Walle, and D. D. Awschalom, Quantum computing with defects, *Proceedings of the National Academy of Sciences* **107**, 8513 (2010).
 - [9] M. W. Doherty, N. B. Manson, P. Delaney, F. Jelezko, J. Wrachtrup, and L. C. L. Hollenberg, The nitrogen-vacancy colour centre in diamond, *Physics Reports* **528**, 1 (2013).
 - [10] S. Castelletto, L. Rosa, and B. C. Johnson, Silicon carbide for novel quantum technology devices, in *Advanced Silicon Carbide Devices and Processing* (InTech, 2015).
 - [11] N. T. Son, C. P. Anderson, A. Bourassa, K. C. Miao, C. Babin, M. Widmann, M. Niethammer, J. U. Hassan, N. Morioka, I. G. Ivanov, F. Kaiser, J. Wrachtrup, and D. D. Awschalom, Developing silicon carbide for quantum spintronics, *Applied Physics Letters* **116**, 190501 (2020).
 - [12] M. E. Bathen and L. Vines, Manipulating single-photon emission from point defects in diamond and silicon carbide, *Advanced Quantum Technologies* **4**, 2100003 (2021).

- [13] M. Atatüre, D. Englund, N. Vamivakas, S.-Y. Lee, and J. Wrachtrup, Material platforms for spin-based photonic quantum technologies, *Nature Reviews Materials* **3**, 38 (2018).
- [14] G. Zhang, Y. Cheng, J.-P. Chou, and A. Gali, Material platforms for defect qubits and single-photon emitters, *Applied Physics Reviews* **7**, 031308 (2020).
- [15] A. M. Deiana, N. Tran, J. Agar, M. Blott, G. D. Guglielmo, J. Duarte, P. Harris, S. Hauck, M. Liu, M. S. Neubauer, J. Ngadiuba, S. Ogreni-Memik, M. Pierini, T. Aarrestad, S. Bahr, J. Becker, A.-S. Berthold, R. J. Bonventre, T. E. M. Bravo, M. Diefenthaler, Z. Dong, N. Fritzsche, A. Gholami, E. Govorkova, K. J. Hazelwood, C. Herwig, B. Khan, S. Kim, T. Klijnsma, Y. Liu, K. H. Lo, T. Nguyen, G. Pezzullo, S. Rasoulnezhad, R. A. Rivera, K. Scholberg, J. Selig, S. Sen, D. Strukov, W. Tang, S. Thais, K. L. Unger, R. Vilalta, B. Krosigk, T. K. Warburton, M. A. Flechas, A. Aportela, T. Calvet, L. Cristella, D. Diaz, C. Doglioni, M. D. Galati, E. E. Khoda, F. Fahim, D. Giri, B. Hawks, D. Hoang, B. Holzman, S.-C. Hsu, S. Jindariani, I. Johnson, R. Kansal, R. Kastner, E. Katsavounidis, J. Krupa, P. Li, S. Madireddy, E. Marx, P. McCormack, A. Meza, J. Mitrevski, M. A. Mohammed, F. Mokhtar, E. Moreno, S. Nagu, R. Narayan, N. Palladino, Z. Que, S. E. Park, S. Ramamoorthy, D. Rankin, S. Rothman, A. Sharma, S. Summers, P. Vischia, J.-R. Vlimant, and O. Weng, Applications and techniques for fast machine learning in science (2021), arXiv:2110.13041 [cs.LG].
- [16] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, *Rev. Mod. Phys.* **91**, 045002 (2019).
- [17] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, Commentary: The materials project: A materials genome approach to accelerating materials innovation, *APL Materials* **1**, 011002 (2013).
- [18] A. Jain, J. Montoya, S. Dwaraknath, N. E. R. Zimmermann, J. Dagdelen, M. Horton, P. Huck, D. Winston, S. Cholia, S. P. Ong, and K. Persson, The materials project: Accelerating materials design through theory-driven data and tools, in *Handbook of Materials Modeling* (Springer International Publishing, 2018).
- [19] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, and C. Wolverton, Materials design and discovery with high-throughput density functional theory: The open quantum materials database (OQMD), *JOM* **65**, 1501 (2013).
- [20] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, and C. Wolverton, The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies, *npj Computational Materials* **1** (2015).
- [21] K. Choudhary, K. F. Garritty, A. C. E. Reid, B. DeCost, A. J. Biacchi, A. R. Hight Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone, A. Davydov, J. Jiang, R. Pachter, G. Cheon, E. Reed, A. Agrawal, X. Qian, V. Sharma, H. Zhuang, S. V. Kalinin, B. G. Sumpter, G. Pilania, P. Acar, S. Mandal, K. Haule, D. Vanderbilt, K. Rabe, and F. Tavazza, The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design, *npj computational materials* **6** (2020).
- [22] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. H. Taylor, L. J. Nelson, G. L. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, and O. Levy, AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations, *Computational Materials Science* **58**, 227 (2012).
- [23] S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, and D. Morgan, AFLOW: An automatic framework for high-throughput materials discovery, *Computational Materials Science* **58**, 218 (2012).
- [24] C. E. Calderon, J. J. Plata, C. Toher, C. Oses, O. Levy, M. Fornari, A. Natan, M. J. Mehl, G. Hart, M. B. Nardelli, and S. Curtarolo, The AFLOW standard for high-throughput materials science calculations, *Computational Materials Science* **108**, 233 (2015).
- [25] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, Universal fragment descriptors for predicting properties of inorganic crystals, *Nature Communications* **8**, 1 (2017).
- [26] L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, and A. Jain, Matminer: An open source toolkit for materials data mining, *Computational Materials Science* **152**, 60 (2018).
- [27] A. M. Ferrenti, N. P. de Leon, J. D. Thompson, and R. J. Cava, Identifying candidate hosts for quantum defects via data mining, *npj Computational Materials* **6**, 1 (2020).
- [28] P. Mehta, M. Bukov, C. Wang, A. G. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, A high-bias, low-variance introduction to machine learning for physicists, *Phys. Rep.* **810**, 1 (2019).
- [29] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (Springer Verlag, Berlin, 2009).
- [30] K. Murphy, *Machine learning : a probabilistic perspective* (MIT Press, 2012).
- [31] Hebnes, O. L., Predicting solid-state qubit material hosts, <https://zenodo.org/record/6345880> (2022).
- [32] G. Kresse and J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, *Physical Review B* **54**, 11169 (1996).
- [33] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized gradient approximation made simple, *Physical Review Letters* **77**, 3865 (1996).
- [34] C. Freysoldt, B. Grabowski, T. Hickel, J. Neugebauer, G. Kresse, A. Janotti, and C. G. Van de Walle, First-principles calculations for point defects in solids, *Reviews of Modern Physics* **86**, 253 (2014).
- [35] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, Python materials genomics (pymatgen): A robust, open-source python library for materials analysis, *Computational Materials Science* **68**, 314 (2013).
- [36] F. Allen, Bergerhoff, and R. Sievers, Crystallographic databases (1987).

- [37] D. Zagorac, H. Müller, S. Ruehl, J. Zagorac, and S. Rehme, Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features, *Journal of Applied Crystallography* **52**, 918 (2019).
- [38] J. O'Mara, B. Meredig, and K. Michel, Materials data infrastructure: A case study of the citrination platform to examine data import, storage, and access, *JOM* (1989) **68**, 2031 (2016).
- [39] O. L. Hebnes, M. E. Bathen, Ø. S. Schøyen, S. G. Winther-Larsen, L. Vines, and M. Hjorth-Jensen, Supplementary Information available at <https://github.com/mhjensen/PredictingSolidStateQubitCandidates> (2022).
- [40] M. Toth and I. Aharonovich, Single photon sources in atomically thin materials, *Annual Review of Physical Chemistry* **70**, 123 (2019).
- [41] W. J. Hardy, C. T. Harris, Y.-H. Su, Y. Chuang, J. Moussa, L. N. Maurer, J.-Y. Li, T.-M. Lu, and D. R. Luhman, Single and double hole quantum dots in strained Ge/SiGe quantum wells, *Nanotechnology* **30**, 215202 (2019).
- [42] J. M. Taylor, P. Cappellaro, L. Childress, L. Jiang, D. Budker, P. R. Hemmer, A. Yacoby, R. Walsworth, and M. D. Lukin, High-sensitivity diamond magnetometer with nanoscale resolution, *Nature Physics* **4**, 810 (2008).
- [43] G. Balasubramanian, P. Neumann, D. Twitchen, M. Markham, R. Kolesov, N. Mizuochi, J. Isoya, J. Achard, J. Beck, J. Tissler, V. Jacques, P. R. Hemmer, F. Jelezko, and J. Wrachtrup, Ultralong spin coherence time in isotopically engineered diamond, *Nature Materials* **8**, 383 (2009).
- [44] P. E. Barclay, K.-M. C. Fu, C. Santori, A. Faraon, and R. G. Beausoleil, Hybrid nanocavity resonant enhancement of color center emission in diamond, *Physical Review X* **1** (2011).
- [45] L. Gordon, J. R. Weber, J. B. Varley, A. Janotti, D. D. Awschalom, and C. G. Van de Walle, Quantum computing with defects, *MRS Bulletin* **38**, 802 (2013).
- [46] L. J. Rogers, K. D. Jahnke, M. H. Metsch, A. Sipahigil, J. M. Binder, T. Teraji, H. Sumiya, J. Isoya, M. D. Lukin, P. Hemmer, and F. Jelezko, All-optical initialization, readout, and coherent preparation of single silicon-vacancy spins in diamond, *Physical Review Letters* **113**, 263602 (2014).
- [47] M. K. Bhaskar, D. D. Sukachev, A. Sipahigil, R. E. Evans, M. J. Burek, C. T. Nguyen, L. J. Rogers, P. Siyushev, M. H. M. abd H. Park, F. Jelezko, M. Loncar, and M. D. Lukin, Quantum nonlinear optics with a germanium-vacancy color center in a nanoscale diamond waveguide, *Phys. Rev. Lett.* **118**, 223603 (2017).
- [48] M. Widmann, S.-Y. Lee, T. Rendler, N. T. Son, H. Fedder, S. Paik, L.-P. Yang, N. Zhao, S. Yang, I. Booker, A. Denisenko, M. Jamali, S. A. Momenzadeh, I. Gerhardt, T. Ohshima, A. Gali, E. Janzén, and J. Wrachtrup, Coherent control of single spins in silicon carbide at room temperature, *Nature Materials* **14**, 164 (2014).
- [49] D. J. Christle, A. L. Falk, P. Andrich, P. V. Klimov, J. Ul Hassan, N. T. Son, E. Janzén, T. Oshima, and D. D. Awschalom, Isolated electron spins in silicon carbide with millisecond coherence times, *Nature Materials* **14**, 160 (2015).
- [50] S. Castelletto, B. C. Johnson, V. Ivády, N. Stavrias, T. Umeda, A. Gali, and T. Oshima, A silicon carbide room-temperature single-photon source, *Nature Materials* **13**, 151 (2014).
- [51] S. A. Zargaleh, S. Hameau, B. Eble, F. Margaihan, H. J. von Bardeleben, J. L. Cantin, and W. Gao, Nitrogen vacancy center in cubic silicon carbide: A promising qubit in the 1.5 μm spectral range for photonic quantum networks, *Physical Review B* **98**, 165203 (2018).
- [52] A. L. Falk, B. B. Buckley, G. Calusine, W. F. Koehl, V. V. Dobrovitski, A. Politi, C. A. Zorman, P. X.-L. Feng, and D. D. Awschalom, Polytype control of spin qubits in silicon carbide, *Nature Communications* **4** (2013).
- [53] J. T. Muhonen, J. P. Dehollain, A. Laucht, F. E. Hudson, R. Kalra, T. Sekiguchi, K. M. Itoh, D. N. Jamieson, J. C. McCallum, A. S. Dzurak, and A. Morello, Storing quantum information for 30 seconds in a nanoelectronic device, *Nature Nanotechnology* **9**, 986 (2014).
- [54] A. Durand, Y. Baron, W. Redjem, T. Herzig, A. Benali, S. Pezzagna, J. Meijer, A. Y. Kuznetsov, J.-M. Gérard, I. Robert-Philip, M. Abbarchi, V. Jacques, G. Cassaboïs, and A. Dréau, Broad diversity of near-infrared single-photon emitters in silicon, *Physical Review Letters* **126**, 083602 (2021).
- [55] W. Redjem, A. Durand, T. Herzig, A. Benali, S. Pezzagna, J. Meijer, A. Y. Kuznetsov, H. S. Nguyen, S. Cuff, J.-M. Gérard, I. Robert-Philip, B. Gil, D. Caliste, P. Pochet, M. Abbarchi, V. Jacques, A. Dréau, and G. Cassaboïs, Single artificial atoms in silicon emitting at telecom wavelengths, *Nature Electronics* **3**, 738 (2020).
- [56] T. T. Tran, C. Elbadawi, D. Totonjian, C. J. Lobo, G. Grosso, H. Moon, D. R. Englund, M. J. Ford, I. Aharonovich, and M. Toth, Robust multicolor single photon emission from point defects in hexagonal boron nitride, *ACS Nano* **10**, 7331 (2016).
- [57] T. T. Tran, K. Bray, M. J. Ford, M. Toth, and I. Aharonovich, Quantum emission from hexagonal boron nitride monolayers, *Nature Nanotechnology* **11**, 37 (2016).
- [58] F. Hayee, L. Yu, J. L. Zhang, C. J. Ciccarino, M. Nguyen, A. F. Marshall, I. Aharonovich, J. Vucković, P. Narang, T. F. Heinz, and J. A. Dionne, Revealing multiple classes of stable quantum emitters in hexagonal boron nitride with correlated optical and electron microscopy, *Nature Materials* **19**, 534 (2020).
- [59] A. J. Morfa, B. C. Gibson, M. Karg, T. J. Karle, A. D. Greentree, P. Mulvaney, and S. Tomljenovic-Hanic, Single-photon emission and quantum characterization of zinc oxide defects, *Nano Letters* **12**, 949 (2012).
- [60] C. Stewart, M. Kianinia, R. Previdi, T. T. Tran, I. Aharonovich, and C. Bradac, Quantum emission from localized defects in zinc sulfide, *Optics Letters* **44**, 4873 (2019).
- [61] H. Bluhm, S. Foletti, I. Neder, M. Rudner, D. Mahalu, V. Umansky, and A. Yacoby, Dephasing time of GaAs electron-spin qubits coupled to a nuclear bath exceeding 200 μs , *Nature Physics* **7**, 109 (2010).
- [62] F. L. Roux, K. Gao, M. Holmes, S. Kako, M. Arita, and Y. Arakawa, Temperature dependence of the single photon emission from interface-fluctuation GaN quantum dots, *Scientific Reports* **7** (2017).
- [63] A. M. Berhane, K.-Y. Jeong, C. Bradac, M. Walsh, D. Englund, M. Toth, and I. Aharonovich, Photophysics of GaN single-photon emitters in the visible spectral range, *Physical Review B* **97** (2018).

- [64] Y. Xue, H. Wang, N. Xie, Q. Yang, F. Xu, B. Shen, J. Jie Shi, D. Jiang, X. Dou, T. Yu, and B. quan Sun, Single-photon emission from point defects in aluminum nitride films, *The Journal of Physical Chemistry Letters* **11**, 2689 (2020).
- [65] A. Inselberg and B. Dimsdale, Parallel coordinates: a tool for visualizing multi-dimensional geometry, in *Proceedings of the First IEEE Conference on Visualization: Visualization 90* (IEEE Comput. Soc. Press, 1990).
- [66] A. Inselberg, The plane with parallel coordinates, *The Visual Computer* **1**, 69 (1985).
- [67] I. T. Jolliffe, *Principal Component Analysis* (Springer, Berlin, 2002).
- [68] C. Beleites and R. Salzer, Assessing and improving the stability of chemometric models in small sample size situations, *Analytical and Bioanalytical Chemistry* **390**, 1261 (2008).
- [69] T. Chen and C. Guestrin, Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16 (Association for Computing Machinery, New York, NY, USA, 2016) p. 785–794.
- [70] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, *npj Computational Materials* **2**, 1 (2016).
- [71] Y. Gao, Y. Wu, Q. Huang, M. Ma, Y. Pan, M. Xiong, Z. Li, Z. Zhao, J. He, and D. Yu, *Journal of Applied Physics* **121**, 225103 (2017).
- [72] Y. Cai, J. Xiong, Y. Liu, and X. Xu, Electronic structure and chemical hydrogen storage of a porous sp^3 tetragonal BC_2N compound, *Journal of Alloys and Compounds* **724**, 229 (2017).
- [73] H. Li, X. Xiao, J. Tie, and J. Lu, Electronic and magnetic properties of bare armchair BC_2N nanoribbons, *Journal of Magnetism and Magnetic Materials* **426**, 641 (2017).
- [74] C.-L. Jiang, W. Zeng, F.-S. Liu, B. Tang, and Q.-J. Liu, The shape type of bonds and the direction of phonons in orthorhombic BC_2N from first-principles calculations, *Journal of Physics and Chemistry of Solids* **140**, 109349 (2020).
- [75] D. Wang, L. Liu, and H. L. Zhuang, Spin qubit based on the nitrogen-vacancy center analog in a diamond-like compound C_3BN , *Journal of Applied Physics* **130**, 225702 (2021).
- [76] J. Liu, K. Konthasinghe, M. Davanço, J. Lawall, V. Anant, V. Verma, R. Mirin, S. W. Nam, J. D. Song, B. Ma, Z. S. Chen, H. Q. Ni, Z. C. Niu, and K. Srinivasan, Single self-assembled InAs/GaAs quantum dots in photonic nanostructures: The role of nanofabrication, *Physical Review Applied* **9** (2018).
- [77] D. Ahn, J. D. Song, S. S. Kang, J. Y. Lim, S. H. Yang, S. Ko, S. H. Park, S. J. Park, D. S. Kim, H. J. Chang, and J. Chang, Intrinsically p-type cuprous iodide semiconductor for hybrid light-emitting diodes, *Scientific Reports* **10** (2020).
- [78] S. R. Zhang, L. H. Xie, S. D. Ouyang, X. W. Chen, and K. H. Song, Electronic structure, chemical bonding and optical properties of the nonlinear optical crystal $ZnGeP_2$ by first-principles calculations, *Physica Scripta* **91**, 015801 (2015).
- [79] G. C. Xing, K. J. Bachmann, J. B. Posthill, and M. L. Timmons, $ZnGeP_2$: A wide bandgap chalcopyrite structure semiconductor for nonlinear optical applications, *MRS Proceedings* **162** (1989).
- [80] L. Caspani, C. Xiong, B. J. Eggleston, D. Bajoni, M. Liscidini, M. Galli, R. Morandotti, and D. J. Moss, Integrated sources of photon quantum states based on nonlinear optics, *Light: Science & Applications* **6**, e17100 (2017).
- [81] G. Scappucci, C. Kloeffer, F. A. Zwanenburg, D. Loss, M. Myronov, J.-J. Zhang, S. De Franceschi, G. Katsaros, and M. Veldhorst, The germanium quantum information route, *Nature reviews. Materials* (2020).
- [82] V. A. Mukhanov, D. Vrel, P. S. Sokolov, Y. Le Godec, and V. L. Solozhenko, Ultra-fast mechanochemical synthesis of boron phosphides, BP and $B_{12}P_2$, *Dalton transactions : an international journal of inorganic chemistry* **45**, 10122 (2016).
- [83] H. Zhang, X. Ma, Q. Lin, Z. Zeng, H. Wang, L. S. Li, H. Shen, Y. Jia, and Z. Du, High-brightness blue InP quantum dot-based electroluminescent devices: The role of shell thickness, *The Journal of Physical Chemistry Letters* **11**, 960 (2020).
- [84] Y.-H. Won, O. Cho, T. Kim, D.-Y. Chung, T. Kim, H. Chung, H. Jang, J. Lee, D. Kim, and E. Jang, Highly efficient and stable InP/ZnSe/ZnS quantum dot light-emitting diodes, *Nature* **575**, 634 (2019).
- [85] J. Rickman, T. Lookman, and S. Kalinin, Materials informatics: From the atomic-level to the continuum, *Acta Materialia* **168**, 473 (2019).
- [86] Y. He, S. Gorman, D. Keith, L. Krantz, J. Keizer, and M. Simmons, A two-qubit gate between phosphorus donor electrons in silicon, *Nature* **571**, 371 (2019).
- [87] G. Thiering and A. Gali, Color centers in diamond for quantum applications, in *Diamond for Quantum Applications* (Elsevier, 2020) pp. 1–36.
- [88] L. Wang, T. Maxisch, and G. Ceder, Oxidation energies of transition metal oxides within the GGA+U framework, *Physical Review B* **73** (2006).
- [89] V. Stevanović, S. Lany, X. Zhang, and A. Zunger, Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies, *Physical Review B* **85** (2012).
- [90] T. Thonhauser, V. R. Cooper, S. Li, A. Puzder, P. Hyldgaard, and D. C. Langreth, Van der Waals density functional: Self-consistent potential and the nature of the van der Waals bond, *Physical Review B* **76** (2007).
- [91] J. Klimeš, D. R. Bowler, and A. Michaelides, Van der Waals density functionals applied to solids, *Physical Review B* **83** (2011).
- [92] F. Tran and P. Blaha, Accurate band gaps of semiconductors and insulators with a semilocal exchange-correlation potential, *Physical Review Letters* **102** (2009).
- [93] K. Choudhary, Q. Zhang, A. C. Reid, S. Chowdhury, N. V. Nguyen, Z. Trautt, M. W. Newrock, F. Y. Congo, and F. Tavazza, Computational screening of high-performance optoelectronic materials using OptB88vdw and TB-mBJ formalisms, *Scientific Data* **5** (2018).
- [94] P.-P. D. Breuck, M. L. Evans, and G.-M. Rignanese, Robust model benchmarking and bias-imbalance in data-driven materials science: a case study on MODNet, *Journal of Physics: Condensed Matter* **33**, 404002 (2021).

ACKNOWLEDGMENTS

The work of LV and MEB was supported by the Research Council of Norway and the University of Oslo through the frontier research projects FUNDAMeNT (no. 251131) and QuTe (no. 325573). The work of MEB was supported by an ETH Zurich Postdoctoral Fellowship. The work of MHJ was supported by the U.S. Department of Energy, Office of Science, office of Nuclear Physics under grant No. DE-SC0021152 and U.S. National Science Foundation Grants No. PHY-1404159 and PHY-2013047. The work of SGWL and ØSS was supported by the Norwegian Directorate for International Cooperation and Quality Enhancement in Higher Education (DIKU) which supports the Center for Computing in Science Education (CCSE).

AUTHOR CONTRIBUTIONS

MEB, LV and MHJ conceived the main theme of the project. OH developed the programs and performed the bulk of the work while MEB lead the writing process. All authors have contributed to the writing of the paper and the discussion and analysis of the data.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to bathen@aps.ee.ethz.ch. Supplementary material is available at <https://github.com/mhjensen/PredictingSolidStateQubitCandidates>.